The Role of Randomization in Inference
Author(s): Dennis V. Lindley
Reviewed work(s):
Source: *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, Vol. 1982, Volume Two: Symposia and Invited Papers (1982), pp. 431-446
Published by: The University of Chicago Press on behalf of the Philosophy of Science Association
Stable URL: http://www.jstor.org/stable/192435
Accessed: 27/02/2012 16:23

The Role of Randomization in Inference

Dennis V. Lindley

1. Introduction

> Who is there that has not longed that the power and privilege of selection among alternatives should be taken away from him in some important crisis of his life, and that his conduct should be arranged for him, either this way or that, by some divine power if it were possible, -- by some patriarchal power in the absence of divinity, -- or by chance even, if nothing better than chance could be found to do it? Anthony Trollope *Phineas Finn* Vol. II, Ch. LX.

In the design and analysis of an experiment there are several places where an element of randomization can be used: the design can be selected at random, the result can have a random element adjoined to it, or the random element already present can be used in the analysis. The first technique is much used by statisticians; for example, in making a survey of a population, Basu (1980) calls it prerandomization. The second, postrandomization, is hardly ever advocated. Statisticians' attitude towards the third is ambivalent; randomization is used when making a significance test but denied when effecting a likelihood analysis. In this paper we discuss all three ideas and argue that randomization is undesirable in each case. This follows easily from the Bayesian approach but also from two, simple principles; of conditionality and similarity. Both viewpoints lead to the likelihood principle which completely denies the relevance of any random element once the data are available. The situation with experimental design is subtler. It is argued that it is important to ensure that the effect observed is really due to the factor that appears to cause it and not to some other, unrecognized factor. To do this a procedure is required which is closely related to, but different from, randomization: we call it haphazardness. Thus the practical advice to randomize is sensible, but to make use of this randomization in the analysis is not. We also discuss the important concept of exchangeability in the use of experimental data.

PSA 1982, Volume 2, pp. 431-446

2. Analysis of Data

The discussion begins with consideration of whether randomization has any role to play in the analysis of data: we shall later discuss the role in the design of an experiment. This is the correct, logical order since one cannot sensibly discuss the choice between two experimental designs until one has discovered how to analyze the data they might yield.

An experiment consists of a protocol to be followed and a specification of the possible outcomes of pursuing the protocol. Denote by X the set of outcomes x. In analyzing the data we have to report the effect the data have on something. We follow the usual statistical device of introducing a space Θ of parameters θ and linking it to X by supposing that for each θ there is a probability distribution over X. Our task is to make a statement about θ on the basis of the result x of the experiment. The form of this statement need not concern us: our results will be valid for any form.

An experiment is a triplet $E = \{X, \Theta, p(x|\theta)\}$ where $p(x|\theta)$ is the probability[1] of x, given θ: that is $p(x|\theta) > 0$ for all (x, θ) and $\sum_{x \varepsilon X} p(x|\theta) = 1$ for all θ. We will follow Basu (1975) in referring to the information[2] (about θ) contained in an experiment E that yields data x: write $\mathcal{I}(E, x)$. We want to see what element randomization might play in the evaluation of this information.

Consider first the Bayesian view. This says that (E, x) provides information about θ by changing the probability distribution for θ. The change is effected by multiplying the original distribution by the likelihood function $p(x|\theta)$, for the observed x, and then normalizing the result. In other words, the information in (E, x) is entirely conveyed by the likelihood function for x. A few words in amplification may be in order.

The function $p(x|\theta)$, considered as a function of x for fixed θ is a probability mass function: we sometimes write $p(\cdot|\theta)$ to emphasize the distinction between the function and its value $p(x|\theta)$ at x. Similarly $p(x|\cdot)$ is the likelihood function of θ for data x. The likelihood function does not sum (over Θ) to 1 and indeed any constant multiple $cp(x|\cdot)$ will serve as a likelihood. This is clear from Bayes theorem. The Bayesian argument therefore immediately leads to the

Likelihood principle $\mathcal{I}(E, x) = \mathcal{I}(\Theta, p(x|\cdot))$.

In words, given x, the only elements of $E = \{X, \Theta, p(\cdot|\cdot)\}$ that are relevant in extracting the evidence are Θ and $p(x|\cdot)$ for the observed x. Notice that the principle enables us to dispense with X and with $p(\cdot|\cdot)$ except for the observed x. In particular no randomization whatsoever enters into the analysis. An alternative expression of the principle is to say that if E_1 and E_2 are two experiments sharing a common Θ and results x_1 and x_2 share a common likelihood, $p(x_1|\cdot) = cp(x_2|\cdot)$, then $\mathcal{I}(E_1, x_1) = \mathcal{I}(E_2, x_2)$, since both are equal to $\mathcal{I}(\Theta, p(x_1|\cdot))$.

Randomization, or more precisely postrandomization, can take two forms. External randomization adds a random element to x. Internal randomization uses the random element already present in $p(\cdot|\theta)$. As an example of the latter consider a significance test of the null hypothesis $\theta = \theta_0$: this uses a significance level

$$\sum_{x\varepsilon R} p(x|\theta_0)$$

where R is a set (the rejection region) in X. The likelihood principle rules out both forms of randomization. We now show that two simple principles, quite outside the Bayesian paradigm, also lead to the likelihood principle. The argument is due to Basu (1975), following earlier work by Birnbaum (1962,1972).

Consider two experiments $E_i = \{X_i, \Theta, p(x_i|\theta)\}$ $(i = 1, 2)$ sharing a common parameter space Θ, and derive from them a new experiment E, called a mixture experiment, with the protocol: with known probability α perform E_1, with complementary probability $(1 - \alpha)$ perform E_2. (It is important that α does not depend on θ.) The result of performing E will be a pair (i, x_i), $i = 1, 2$.

Conditionality principle $\mathcal{I}(E, (i, x_i)) = \mathcal{I}(E_i, x_i)$.

In words, this says that if, in performing E, you perform E_1 and obtain x_1, it does not matter at all that you might have performed E_2, and similarly with the roles of E_1 and E_2 reversed. The principle is discussed below.

The experiments E_1 and E_2 above are said to be similar if there is a one-to-one correspondence between the elements of X_1 and X_2 which preserves the probability distributions for all θ. If the correspondence is $x_2 = gx_1$ or $x_1 = g^{-1}x_2$, then $p(x_2|\theta) = p(x_1|\theta)$ whenever $x_2 = gx_1$.

Similarity principle If E_1 and E_2 are similar then

$$\mathcal{I}(E_1, x_1) = \mathcal{I}(E_2, x_2) \text{ whenever } x_2 = gx_1.$$

We now prove that these two principles together imply the likelihood principle. Let E_1 and E_2 be any two experiments with a common Θ and with x_1, x_2 such that $p(x_1|\theta) = cp(x_2|\theta)$ for all θ and some c: that is, x_1 and x_2 yield the same likelihood function. Let E be the mixture experiment derived from E_1 and E_2 with $\alpha = 1/(1 + c)$ and $1 - \alpha = c/(1 + c)$. Then by the conditionality principle

$$\left.\begin{aligned} &I(E, (1, x_1)) = I(E_1, x_1) \\ \text{and}\quad &I(E, (2, x_2)) = I(E_2, x_2) \end{aligned}\right\} \tag{2.1}$$

The likelihood for x_1 in E is $p(x_1|\cdot)/(1 + c)$ and for x_2 is $cp(x_2|\cdot)/(1 + c)$, but $p(x_1|\cdot) = cp(x_2|\cdot)$, so that x_1 and x_2 in E have the same likelihood; indeed, in E, $p(x_1|\cdot) = p(x_2|\cdot)$. Now consider a transformation of E into itself with $gx_1 = x_2$ and $gx_2 = x_1$ but otherwise $gx = x$. By the similarity principle

$$I(E, (1, x_1)) = I(E, (2, x_2)). \tag{2.2}$$

Combining (2.1) and (2.2) we have $I(E_1, x_1) = I(E_2, x_2)$ which is the likelihood principle.

It has therefore been shown that the conditionality and similarity principles together rule out all forms of randomization in the analysis of data. Let us therefore consider the two principles. The similarity one is so obviously correct that it is superfluous to comment. The conditionality principle is perhaps not so transparent. Statisticians' attitudes to it have been ambivalent. Sometimes it is accepted as obvious, as when a random sample of size n is taken, call this E_n, and the analysis ignores other possible E_m, $m \neq n$. Sometimes it is vigorously resisted, as with significance tests when the likelihood principle it implies is violated. The ambivalence is most clearly seen in connection with ancillary statistics.

A function, or statistic, t(x) is <u>ancillary</u> if its distribution does not depend on θ. We may then write

$$p(x|\theta) = p(t(x))p(x|t(x), \theta)$$

and E is clearly a mixture of experiments E_t, E_t being the experiment in which the ancillary takes the value t. Statisticians often, to reduce E to E_t, use an ancillary but then go on to use a significance test based on a region in the sample space of E_t, thus violating the likelihood principle. Cox (1980) is an example. Embarrassment is caused by the fact that there may exist several ancillaries and the unresolved question then arises of which one to use.

My own attitude to the conditionality principle is that, on reflection, it is entirely acceptable. The fact that, in performing E_1, you might have performed E_2, where the choice between E_1 and E_2 was by a chance mechanism known to you, seems irrelevant. There is the additional fact that it leads to the likelihood principle, which is not known to give rise to difficulties.

One objection to our argument says that sometimes the basic structure of E as a triplet $\{X, \Theta, p(x|\theta)\}$ is inappropriate because no natural parameter space Θ exists. These are situations in which there is one natural probability measure over X and the purpose of the experiment is to see whether this probability reasonably obtains. The description of E is a pair $\{X, p(x)\}$ and $p(x)$ is often called the null hypothesis. A significance test of the null hypothesis is obtained, as above, by selecting a rejection region R of X with level

$$\sum_{x\varepsilon R} p(x) = \alpha.$$

In this context the likelihood principle is vacuous because there is no likelihood. The conditionality and similarity principles still make sense, with the set Θ containing a single point, and the proof given above says that $E_1 = \{X_1, p_1(x)\}$ and $E_2 = \{X_2, p_2(x)\}$ have the same information, about whether their respective null hypotheses obtain, if $p_1(x_1) = p_2(x_2)$. This condition does not typically obtain with a significance test.

However, the conditionality principle is doubtful in this context since queries about $p_1(x)$ are not necessarily the same as those about $p_2(x)$ or about the mixture probability $\alpha p_1(x) + (1 - \alpha)p_2(x)$. In the treatment with a nondegenerate parameter space the shared value of θ means that the query concerns, and the information is about, this common feature. This is not true in the case of only one probability distribution.

My response to this line of reasoning is that it does not make sense to test a hypothesis without alternatives in mind. Data x may be astonishing according to the null hypothesis (or p(x) is small) but, unless we can think of an alternative with appreciably larger probability for x, should we not still accept the null; for rare events should happen sometimes, albeit rarely. Any consideration of what constitutes a "good" significance test involves tacit discussion of alternatives. Furthermore, by the likelihood principle, we know that, whatever alternatives were considered, only p(x) from the null value (plus $p(x|\cdot)$ from the alternatives) would be required, so why use p(x') for $x' \neq x$ when the alternatives are unspecified? Another point is that no one has so far succeeded in constructing a satisfactory theory of significance tests of a null hypothesis without introducing alternatives.

3. Design of experiments

We now turn to the question of whether randomization plays any role in the design of an experiment, as distinct from its analysis. Let us first note that from a Bayesian view it does not. In that view the design of an experiment consists of a choice between a number of decisions, each one being a decision to carry out a particular experiment. Any decision has its expected utility, as seen by the experimenter, and the best decision is the one that maximizes this value. Then if two possible experiments (or decisions) E_1 and E_2 have respectively expected utilities u_1 and u_2, a random choice between them, in which E_1 is selected with probability p, has expected utility $pu_1 + (1 - p)u_2$. If $u_1 \neq u_2$, and for definiteness $u_1 > u_2$, this value is less than u_1 and the random choice is never preferred. It is only when $u_1 = u_2$ that the randomized experiment need be considered but then its expected utility is the same as the common value for E_1 and E_2 and therefore has no advantage over them. The extension from two to any number of experiments is immediate. In summary, randomization can do no good, and may do harm in the decrease of utility.

The conditionality principle also denies the value of experimental randomization because it says that if we randomly choose between E_1 and E_2 and in fact perform E_1, then the fact that E_1 was randomly selected is irrelevant in the subsequent analysis.

Consequently there is a direct conflict between the Bayesian view, the conditionality principle, and accepted practical wisdom which says that randomization is an essential part of good experimental design and that its introduction by Fisher was a major, scientific advance. In the rest of this paper we resolve this apparent conflict by showing that it is not the use of a randomizing device that is important but rather a certain haphazard quality that is needed and this haphazardness is accommodated in the Bayesian approach. Thus a minor change of emphasis confirms the practical wisdom and supports Fisher's brilliance whilst preserving the logical, coherent view of inference.

The conflict resolution is most easily discussed in the context of an example. Suppose that there is an even number, 2n, of units available for experimentation. A unit may be a person in a medical experiment or a plot of land in an agricultural field trial. Half of the units are to be given a treatment T, the other half are to be denied the treatment $\bar{T}$: in medical language they will receive a placebo. The only unresolved question in the experimental design is which n units are to be given the treatment. One possibility is to assign the treatment at random so that each of the $(2n)!/(n!)^2$ possibilities has the same chance of being selected. Some time after the treatment has, or has not, been applied to a unit a measurement Y is to be made on the unit, the possible values of Y being conventionally 1 or 0. In the medical situation these may be "dead" or "alive": in the field trial "infected, or not" The purpose of the experiment is to judge whether or not the treatment is beneficial (in the sense of reducing deaths or infections) and specifically whether it should be given to other units similar to those units that took part in the experiment.

Before proceeding with the example there is one important matter to be considered. Suppose that some of the units can be recognized[3] as being different from the others in a way that might reasonably affect Y. Thus we might recognize the females in the medical experiment as being less likely to die than the males. In that case the common view amongst practitioners, and a view which can be supported by arguments within the Bayesian framework, is that randomization should *not* take place over all 2n units but only within similar units: in the example, within the women, and within the men. We speak of sex here as being a covariate: or we say that the experiment is divided into blocks, here two, the males and females. It will therefore be supposed that no such relevant subsets can be recognized amongst the 2n units: relevant, that is for Y, and recognized by the experimenter. It is then that randomization becomes a serious possibility and, as we next show, tries to guard against relevant subsets that are *not* recognized.

To illustrate the possibilities suppose that the treatment is not assigned at random but according to some rule or some deliberate policy. Suppose moreover that the assignment depends in an *unrecognized* way on some quantity X, which also assumes only values 1 and 0, and that X affects Y: in the language above, X is a covariate. The key point here is that X is not recognized but is relevant to T and Y. Since a randomizing device is not associated with any quantity X (for this is part of what we mean by "random") the assignment of the treatment at random tends to avoid the association with X.

The following example taken from Lindley and Novick (1981) shows what might happen. Table 1 gives the result of an experiment on

TABLE 1

Effect of a treatment, T, on death rates

	Y = 1, dead	Y = 0, alive		Death rate
T	20	20	40	50%
$\bar{T}$	24	16	40	60%

80(n = 40) persons recording Y = 1, dead, and Y = 0, alive. The death rate for the treated individuals at 50% is lower than that for those receiving the placebo at 60%. The conclusion, ignoring sampling error, is that the treatment has been efficacious in reducing the death rate by 10%. However, it happened that an unrecognized covariate was present and Table 2 shows the results that were obtained after the experiment has been completed and the covariate was subsequently recognized. Notice that the covariate is associated with the treatment because 30 treated patients had X = 1, whereas only 10 had X = 0; the numbers being reversed for those having the placebo. The covariate is also associated with Y, the death rates for patients with X = 1 being much lower than for those with X = 0. We thus have the situation described above with the unrecognized covariate being associated with both treatment and response, Y. Now look at the death rates within the previously unrecognized classes, X = 1 and X = 0. Within *both* of these the effect of the treatment is to *increase* the death rate by 10%, exactly the amount by which the treatment was seen to *decrease* the death rate in Table 1 before the covariate was recognized. The paradox that a treatment can be bad for men (X = 1) and bad for women (X = 0) (Table 2) but good for people (Table 1) is usually known as Simpson's paradox (1951) though it is discussed in Cohen and Nagel (1934).

TABLE 2

Effect of a treatment, T, on death rates,
allowing for a covariate X.

X = 1	Y = 1, dead	Y = 0, alive		Death rate
T	12	18	30	40%
$\bar{T}$	3	7	10	30%
X = 0				
T	8	2	10	80%
$\bar{T}$	21	9	30	70%

The example is enough to demonstrate the need to ensure that the allocation of treatments to units is done in a way that is unaffected by some covariates, namely those that might affect Y, for otherwise differences apparently attributable to treatment might really be due to the covariate. In our example the final effect is exactly the opposite of the apparent one. It is this danger that randomization is designed to avoid. Before we see how well it performs, let us be more precise about what is meant by a randomizing device.

The archetype randomization tool is a table of random numbers. A table of numbers is random for you if the chance of any digit in any place is one tenth irrespective of the digits in any other places: indeed, irrespective of anything else. Consequently the use of a table of random numbers would appear to ensure that the treatment allocation is unaffected by a covariate. However it is not as simple as that. In a finite set of 2n units there is a chance, albeit a very small one, that all the treated units have $X = 1$ and all the untreated $X = 0$. This chance may increase substantially when it is remembered that the number of possible covariates is very large. It is therefore prudent, after having allocated the treatments to units by randomization but before conducting the experiment, to inspect the allocation to see whether it suggests a covariate with which it might be associated. Our understanding is that this inspection, after randomization but before experimentation, is carried out by practitioners; that they would dismiss an allocation that appeared to them unsatisfactory and go on to perform a second randomization. Thus if the allocation of treatment to males in our medical example would have resulted in all the treated males being of one blood type and the untreated ones being of the opposite type, and if it is thought that blood type might affect the response Y, even though this had not been thought of before the allocation suggested the idea, a new allocation would be selected.

If this inspection of the results of randomization is admitted to be a useful precaution, one might ask why randomize in the first place? Why not take a possible allocation, consider it for the presence of possible covariates, as with the randomized allocation, and accept it as the design? We describe a possible allocation that the experimenter judges to be free of covariate interference to be haphazard. Randomization may be a convenient way of producing a haphazard design. We argue that it is the haphazard nature, and not the randomization, that is important. It was a major scientific advance when Fisher recognized this need for precaution in design.

Let us now return to the Bayesian position. Therein it is necessary to consider each possible experiment, that is, each possible allocation, and evaluate its expected utility. If an allocation has a possible covariate, perhaps recognized, like the blood-type above, only after considering that design, then it is necessary to assess its likely effect in order to evaluate the utility. This would be complicated and hardly likely to be worth the effort if the effect is small, as it will be since covariates with large effects will almost certainly have been recognized at an earlier stage in the design. It seems therefore that a reasonable approximation to the optimum design would be to select a haphazard design in the sense of haphazard used above, namely unlikely to involve a relevant covariate.[4]

Consequently the two, apparently conflicting, views of the randomizer and the Bayesian have been brought into agreement. It is the haphazard nature of the allocations, not the random element, that is important; and the use of a haphazard design saves the Bayesian a lot of trouble, with small chance of any appreciable gain, by producing a situation relatively easy to analyze. A further point is that a detailed, Bayesian consideration of possible covariates would almost certainly not be robust in that the analysis might be sensitive to small changes in the judgments about covariates.

4. Use of Experimental Results

There is another aspect of experimental design that is worth exploring because it helps understand how we may make inductive inferences. The basic purpose of an experiment is to help our understanding of the world in order to make future judgments or actions more wisely than otherwise. Thus, in the medical example we wish to know whether to apply the treatment to other people who did not take part in the trial -- should you receive the treatment? In the agricultural situation a farmer needs advice on which variety to plant. As soon as this is recognized we see that it is necessary to establish some connection between experiment and future judgments: between you and the patients in the trial. How is this to be done? Clearly the nature of the connection will depend on the allocation of treatments in the experiment, for whether or not you receive the treatment will depend on considerations that are different from those in the experiment: thus, you will not receive the treatment at random, or even haphazardly. A method for establishing the connection has been suggested by de Finetti (1970) under the name of exchangeability and we now study the way in which the notion can be used in the interpretation of an experiment.

We are concerned with an infinite[5] sequence of random quantities X_i $(i = 1, 2, \ldots)$ each of which can take the values[6] 0 or 1. Consider any set of n of the quantities and the probability that a designated subset of r of them take the value 1, the remaining $(n - r)$ being 0. The infinite sequence is said to be <u>exchangeable</u> if this probability depends only on r and n; so that it does not depend on which n were selected nor on which r of them had the value 1. The probability is written $p(r, n)$. The idea is that it does not matter which particular X's are being considered, any X_i can be exchanged for any X_j. The language may usefully be extended so that we speak of X_{n+1} being exchangeable with $(X_1, X_2, \ldots, X_n)$ if the extended sequence $(X_1, X_2, \ldots, X_n, X_{n+1})$ is exchangeable. To anticipate: are you (X_{n+1}) exchangeable with the patients in the medical trial?

One way in which a sequence can be exchangeable is for each X_i to have the same probability[7], θ say, of being 1, and for all the X's to be independent. Such a sequence is termed Bernoulli. Then $p(r, n) = \theta^r(1 - \theta)^{n-r}$. Furthermore, when this happens $\lim_{n\to\infty} r/n = \theta$, with probability one. De Finetti proved a remarkable theorem that every exchangeable sequence is a mixture of such sequences. Precisely, the probability structure for any exchangeable sequence can be obtained by specifying a probability distribution for θ, the limiting frequency, and requiring the conditional distribution, given θ, to be Bernoulli. We shall refer to θ as the <u>chance</u>[8] that $X_i = 1$.

It is easy to see how these ideas can be applied to experiments. Consider first an experiment, even simpler than the one discussed above, in which n units have been treated similarly with results $Y_1, Y_2, \ldots, Y_n$. A possible interpretation of "similarly" in the last sentence is to make the judgment of exchangeability of the Y's so that the probability of these results is simply p(r, n), where exactly r Y's equal 1, and p(r, n) has the de Finetti form. Consider now a further unit, (n + 1); you may judge that the, as yet unobserved, value for it, Y_{n+1}, is exchangeable with the previous Y's. The uncertainty about that unit may be described by the probability of Y_{n+1}, given $Y^{(n)} = (Y_1, Y_2, \ldots, Y_n)$. Now

$$p\left(Y_{n+1} = 1 \mid Y^{(n)}\right) = p(r + 1, n + 1)/p(r, n)$$

and both the numerator and denominator have the de Finetti form. If the judgment of exchangeability is appropriate this solves the old problem: if an event has been observed to happen r times out of n, what is the probability that it will happen on the next, (n + 1)st, trial?

It is an aside to our main thrust but notice that the answer to this question depends on the probability assigned to the limiting frequency, denoted θ above. To illustrate, suppose that the 20 births last week in a maternity unit had all been male (r = n = 20) then my probability that the next birth will be a boy is a little over 0.51. In contrast if 20 tosses of a thumb tack had all fallen with the point uppermost then my probability that the next toss will have the point uppermost is nearly 1. The reason is that the limiting frequency of male births is known to be stable around 51% whereas no comparable results about thumb tacks are known to me. Previous solutions to this problem have failed to differentiate between sex and thumb tacks.

Now let us return to the original experiment with 2n units, one half receiving a treatment T and the other half a placebo $\bar{T}$. Let us denote the results by $Y^{(2n)}$ when the treatment applications were $T^{(2n)}$. Now consider T_{2n+1}, the application to a new unit, which may be either T or $\bar{T}$. The uncertainty concerns Y_{2n+1} and the relevant probability is

$$p\left(Y_{2n+1} \mid T_{2n+1}, Y^{(2n)}, T^{(2n)}\right)$$
$$= p(Y^{(2n+1)} \mid T^{(2n+1)})/p(Y^{(2n)} \mid T^{(2n)}). \qquad (4.1)$$

The conditions are ready for an assumption of exchangeability. First consider the 2n units that took part in the experiment. These are not exchangeable in the response variable Y because some received the treatment and some did not, but if we take the treated and untreated units separately it is typically satisfactory to suppose each set exchangeable. Thus if units 1 thru n received T and (n + 1) thru 2n $\bar{T}$, each of those subsets could be exchangeable[9] in Y: $Y_1, Y_2, \ldots, Y_n$ and $Y_{n+1}, Y_{n+2}, \ldots, Y_{2n}$. We then call into play the de Finetti representation using chances θ_1 for the first, treated set and θ_2 for the remainder receiving the placebo.[10] A joint probability distribution for θ_1 and θ_2 enables the denominator in (4.1) to be calculated as $p(r_1, n; r_2, n)$ where $r_1(r_2)$ of the treated (untreated) units had $Y = 1$.

The numerator of (1) is not nearly so straightforward. For definiteness let us suppose $T_{2n+1} = T$; that is, the new unit receives the treatment. Then a possibility is to suppose unit (2n + 1) exchangeable with the n units that received the treatment. Were it to receive the placebo, exchangeability with the other n units might be considered and it is the comparison of these two possibilities, for $T_{2n+1} = T$ and $T_{2n+1} = \bar{T}$, that is vital to you as unit (2n + 1). However the factors that govern the choice of T_{2n+1} are clearly rather different from those that affected $T^{(2n)}$. You are not taking part in a designed experiment but are making a reasoned choice in the light of the results of that experiment. To appreciate the difficulties suppose, as above, that associated with each unit is a binary variable X which is not observed and where the allocation of treatment to the units depends on the values of X. Suppose all X_s $(1 \leq s \leq 2n)$ are judged exchangeable and that the induced chance that $X_s = i$ is α_i. We abbreviate this to say $X = i$ has a chance α_i. Suppose $T = j$, given $X = i$, has chance β_{ji}; and finally $Y = k$, given $X = i$ and $T = j$, has chance γ_{kji}. The probability for any of the 2n units that $Y = k$, given $T = j$, is, for known chances

$$\begin{aligned} p(Y = k \mid T = j) &= \sum_i p(Y = k \mid T = j, X = i)p(X = i \mid T = j) \\ &= \sum_i \gamma_{kji}p(T = j \mid X = i)p(X = i)/p(T = j) \\ &= \sum_i \gamma_{kji}\beta_{ji}\alpha_i/\sum_l \beta_{jl}\alpha_l. \end{aligned} \tag{4.2}$$

Next consider the new unit, (2n + 1):

$$p(Y = k \mid T = j) = \sum_i \gamma_{kji}p(X = i \mid T = j) \tag{4.3}$$

but you cannot reasonably infer that $p(X = i \mid T = j)$ is as before since the selection of T_{2n+1} will be based on quite different procedures from those adopted in the experiment. For example, it might be reasonable to suppose T_{2n+1} is independent of X_{2n+1}, then (4.3) becomes

$$p(Y = k \mid T = j) = \sum_i \gamma_{kji}\alpha_i \qquad (4.4)$$

and hence Y_{2n+1} and $Y^{(2n)}$ (equation (4.2)) are governed by different chances and exchangeability is unavailable. Notice that if the allocation of treatment does not depend on X_i (that is, β_{ji} does not involve i) then (4.2) and (4.4) are equal and the usual analysis proceeds even though Y depends on X. Equally if Y does not depend on X (the null case of no treatment effect, $\gamma_{kji} = \gamma_{kj}$) (4.2) and (4.4) are equal.

We therefore see, in amplification of the point made earlier, that if the allocation of treatments in the experiment depends on a hidden covariate that affects the response it will not easily be possible to infer the effect of the treatment on a further unit; in particular, exchangeability will not be a reasonable option. Inference will be possible by placing probability distributions[11] on the sets of chances $\{\alpha_i, \beta_{ji}, \gamma_{kji}\}$, updating them in the light of the data and hence making a probability statement about the chances in (4.4). As we said before, this will be difficult because of the complexity of the analysis and the possible sensitivity of the final result to the assumptions made about the chances.

If the treatment allocation is uninfluenced by the hidden covariate X - in the notation β_{ji} does not depend on i - then the assumption of exchangeability is reasonable and the analysis is straightforward. As we have seen, this can be achieved by a haphazard assignment. Both numerator and denominator in (4.1) can be calculated easily. Thus if $r_1(r_2)$ of the n treated (untreated) responded with $Y = 1$ then the probability that the new unit will have $Y_{2n+1} = 1$ were it to receive the treatment is, from (4.1),

$$p(r_1 + 1, n + 1; r_2, n)/p(r_1, n; r_2, n),$$

whereas were it to receive the placebo, the probability is

$$p(r_1, n; r_2 + 1, n + 1)/p(r_1, n; r_2, n).$$

These probabilities may be found from the exchangeability considerations and de Finetti's theorem.

We conclude by noting two points that are so important that they must be mentioned, yet have so many ramifications that they cannot be discussed adequately at the end of a paper. First, the judgment of exchangeability is a personal one, so that the analyses of the experimental results by two people may differ. In cases of dispute only extended experimentation can bring people into agreement. This is a basic principle in science. The second point, related to this, is that no notion of cause and effect has been introduced into our analysis. The notion that the treatment causes an increase in response is language that is ambiguous and yet totally avoided by the unambiguous judgment of exchangeability. The fact that some people still do not believe that smoking causes lung cancer is a reflection of the fact that they have not felt able to make some exchangeability judgments that they might have been able to make with experiments in which the relationship between smoking and lung cancer was more clear-cut than it is in the presently available data. Causation is therefore a personal matter, not an objective fact, and the recognition of this is an important aid in understanding the nature of the phenomenon we refer to as causation.

Notes

[1] For simplicity in presentation it will be supposed that X and Θ are both finite.

[2] Birnbaum (1962) used the term "evidence".

[3] The word is Fisher's: he spoke of the importance of recognizable subsets.

[4] This use of randomization as an approximate, simplifying technique has been discussed in detail by Rubin (1978).

[5] All the definitions below are for finite sequences but of arbitrarily large size. With a finite upper bound the results would only be approximately true.

[6] The argument extends to more general cases: we take the case of two values only for simplicity in exposition.

[7] This is immediate from the definition when $r = n = 1$.

[8] Some writers use the term propensity.

[9] A slightly stronger assumption is actually needed for what follows but details are omitted.

[10] Some writers like to concentrate on θ_1 and θ_2 to the exclusion of the future unit (2n + 1) and, in particular, to discuss whether the treatment is effective by considering if $\theta_1 = \theta_2$. Our approach makes better sense.

[11] Analogous to a probability for θ in the discussion of exchangeability above.

References

Basu, D. (1975). "Statistical Information and Likelihood." Sankhyā A 37: 1-55, with discussion 56-71.

-------. (1980). "Randomization Analysis of Experimental Data: the Fisher Randomization Test." Journal of American Statistical Association 75: 575-582, with discussion 582-595.

Birnbaum, A. (1962). "On the Foundations of Statistical Inference." Journal of American Statistical Association 57: 269-306, with discussion 307-326.

-----------. (1972). "More on Concepts of Statistical Evidence." Journal of American Statistical Association 67: 858-861.

Cohen, M.R. and Nagel, E. (1934). An Introduction to Logic and Scientific Method. New York: Harcourt Brace.

Cox, D.R. (1980). "Local Ancillarity." Biometrika 67: 279-286.

de Finetti, B. (1970). Teoria delle probabilità. (2 vols.) Torino: G. Einaudi. (As reprinted as Theory of Probability. (2 vols.) (trans.) A. Machí and A. Smith. London: John Wiley & Sons, 1974, 1975.)

Lindley, D.V. and Novick, M.R. (1981). "The Role of Exchangeability in Inference." The Annals of Statistics 9: 45-58.

Rubin, Donald B. (1978). "Bayesian Inference for Causal Effects: The Role of Randomization." The Annals of Statistics 6: 34-58.

Simpson, E.H. (1951). "The Interpretation of Interaction in Contingency Tables." Journal of Royal Statistical Society B 13: 238-241.