Randomization Analysis of Experimental Data: The Fisher Randomization Test

D. BASU*

R.A. Fisher's classic text on the design of experiments is the principal source of inspiration for a mode of data interpretation that is usually characterized as randomization analysis. In Chapter III of this text, Fisher briefly commented on how to make a randomization test on some data generated by a Darwin experiment. Two variants of this randomization test are discussed in this article. The variant that is discussed in Section 4 may be regarded as the forerunner of all nonparametric tests. The original variant of the test is discussed in Section 6. The author concludes that the Fisher randomization test is not logically viable.

KEY WORDS: Prerandomization; Postrandomization; Sufficiency principle; Level of significance; Reference set.

1. INTRODUCTION

Randomization is widely recognized as a basic principle of statistical experimentation. Yet we find no satisfactory answer to the question, Why randomize? In a previous paper (Basu 1978) the question was examined from the point of view of survey statistics. In this article we take an uninhibited frontal view of a part of the randomization methodology generally known as the Fisher randomization test.

R.A. Fisher's classic text *The Design of Experiments* (*DE*) is the principal source of inspiration for a mode of data interpretation that may be characterized as randomization analysis of data. In Chapter III of *DE*, while discussing Galton's analysis of a Darwin experiment with 15 pairs of self-fertilized and cross-fertilized seeds, Fisher cursorily mentioned how one can take advantage of the physical act of randomization to make a test of significance that needs no assumption of normality for the error terms. This idea of Fisher's was immediately generalized by Pitman (1937) and then pushed to its natural boundary by Kempthorne (1952) and many others. Two variants of the Fisher randomization test are discussed in this article. The variant that is discussed in Section 4 may be regarded as the forerunner of all nonparametric tests. The original variant that is discussed in Section 6 may be regarded as one of the two supporting pillars (the other one being the famous case of the "lady tasting tea") of the complex theory of randomization analysis of experimental data. In between the two sections, I have inserted a section entitled: "Did Fisher Change His Mind?" I speculate that in 1956 Fisher had lost a great deal of his early enthusiasm for randomization analysis.

Whether Fisher changed his mind is not the present issue. What I am asking is whether, in the specific instances discussed in this article, it makes sense to compute a significance level (P value) in the manner of the Fisher randomization test. Can any evidential meaning be attached to a P value so computed?

Let us postpone the debate on significance testing in general and nonparametric tests in particular. Let us keep the issue sharply in focus and ask, Can the Fisher randomization test pass the test of common sense?

2. RANDOMIZATION

Let us define randomization as the incorporation of a fully controlled bit of randomness in the process of data generation. Randomization is usually carried out in the manner of items 1 and 2.

1. *Prerandomization.* This is the most common form of randomization. As the name suggests, the data-generation process begins with a fully controlled randomization exercise that determines the actual experimental (or observational) layout. Typical examples are random allocation of treatments in experimental designs and random selection of units in survey sampling. Along with replication and local control (blocking), prerandomization was characterized by Fisher (1960) as one of the three basic principles of statistical experimentation.

2. *Postrandomization.* Abraham Wald (1950) was one of the earliest to consider this kind of randomization as a statistical tool. After data x has been obtained, postrandomization is the generation of a further random entity y whose randomness characteristics may depend on x but are completely known to the randomizer. The statistician's conclusions or decisions are then based on the extended data (x, y). The average performance characteristics of a postrandomized decision rule δ are evaluated by taking into account all possible values of (x, y). With postrandomization, the statistician has a wider choice of attainable risk functions.

3. *Unrecorded randomization.* Occasionally, randomization is allowed to enter into the experimental process in a form quite different from the forms 1 and 2 discussed.

* D. Basu is Professor, Department of Statistics, Florida State University, Tallahassee, FL 32306. This article is based on an invited talk given at the SREB Summer Research Conference in Statistics at Arkadelphia, Arkansas, on June 15, 1978. This research was supported by National Science Foundation Grant MCS 79-04693.

© Journal of the American Statistical Association
September 1980, Volume 75, Number 371
Theory and Methods Section

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_28, © Springer Science+Business Media, LLC 2011

For instance, in a randomized-response survey the subjects may be instructed to respond to the question "Did you truthfully report your gross income in your 1977 tax return?" in the following manner. Each subject tosses a supposedly unbiased coin twice and then answers the question with a "Yes" if the coin yields two heads, with a "No" if the coin yields two tails, or with a truthful "Yes" or "No" if the coin yields a head and a tail. In this data-generation process the statistician may prerandomize to choose his or her sample subjects but has no control over the response randomizations done by the subjects. The statistician can only speculate about the outcomes of the response randomizations but cannot observe them. It may be argued that response randomization need not be classified as a form of experimental randomization. We shall not discuss this kind of randomization in this article. Warner (1965) proposed this kind of survey technique for eliminating evasive-answer bias.

3. TWO FISHER PRINCIPLES

As we said in the introduction, our primary concern is the so-called randomization analysis of data generated by a statistical experiment that has a large measure of prerandomization incorporated in it. It will, however, be useful to clear the deck with a short discussion of postrandomization and the two sides of the sufficiency principle.

Postrandomization injects into the data an element whose randomness characteristics are fully controlled by the experimenter. Let x be the initial data (sample) and let y be the postrandomized variable whose probability distribution, given x, depends only on x. In terms of the extended sample (x, y), the statistic x is sufficient and, as Fisher would put it, summarizes in itself the whole of the relevant information available in (x, y). To incorporate y in the inference-making process will be a violation of

The sufficiency principle: If T is a sufficient statistic then any conclusion that can be validly drawn from a statistical analysis of the data ought to depend on the data only through the statistic T.

In accordance with the sufficiency principle the data should be reduced to the minimal sufficient statistic. Not to reduce the data to the minimal sufficient statistic is to keep open the possibility of being influenced by irrelevant data characteristics such as, say, a postrandomization variable. In this connection it is interesting to read Fisher's (1956, pp. 96–98) comments on a postrandomization test proposed by Bartlett.

According to Fisher, a principal difference between the deductive and the inductive modes of inference is that in the former case valid conclusions (theorems) can be drawn from a partial use of the data (the primary postulates), whereas in the latter case no conclusion can be validly drawn from an examination of only a part of the relevant information core of the data. Fisher was quite concerned with the fact that the maximum likelihood estimator is not always a sufficient statistic. This led him to the conditionality principle and the celebrated recovery-of-ancillary-information method. The Fisher concern about using the whole of the relevant information in the data may be loosely stated as

The insufficiency principle: If the statistic T_1 is not sufficient then an inference-making procedure that depends on the data only through T_1 is insufficient, that is, lacking in substance.

It is not at all surprising, therefore, that Fisher took a rather dim view of nonparametric methods, especially those that make use of only the rank-order statistics. We shall revert to this theme with a Fisher quotation in Section 5.

4. THE FISHER RANDOMIZATION TEST

In Chapter III (Sec. 21) of *DE*, Fisher introduced his randomization test in the following terms: "In these discussions it seems to have escaped recognition that the physical act of randomization, which, as has been shown, is necessary for the validity of any test of significance, affords the means, in respect of any particular body of data, of examining the wider hypothesis in which no normality of distribution is implied." Fisher then gave a brief description of his randomization test as an alternative to the Student's t test. In this section we consider a popular variant of the test that may be regarded as the original permutation test. This is how the test is described in Kempthorne and Folks (1971, p. 342).

Let $x_1, x_2, \ldots, x_n$ be n independent observations on a random variable x. The problem is to test the null hypothesis H_0 that $E(x) = 0$. Under the parametric model that x is normally distributed, the test is usually carried out in terms of the studentized sum $T = \sum x_i$. Under the wider hypothesis (nonparametric model) that the distribution of x is continuous and is symmetric about its mean, the null hypothesis H_0 may be tested in terms of the criterion $T = \sum x_i$ as follows:

Write $\delta_i = \operatorname{sgn} x_i$, $i = 1, 2, \ldots, n$; that is, δ_i is -1 or 1 according as x_i is negative or positive. Note that

$$T = \sum x_i = \sum |x_i| \, \delta_i$$

and that the sample $(x_1, x_2, \ldots, x_n)$ may be split into the two parts

$$(|x_1|, |x_2|, \ldots, |x_n|) \quad \text{and} \quad (\delta_1, \delta_2, \ldots, \delta_n) .$$

Making the standard pretense that we are dealing with random variables and not particular observations, we recognize at once that the $|x_i|$'s are iid and that so also are the δ_i's. Under the null hypothesis, the two parts of the sample are stochastically independent and each δ_i is uniformly distributed over the two-point set $\{-1, 1\}$.

The distribution of the test criterion T is not well defined even under the null hypothesis. If we fix the $|x_i|$'s at their observed values and regard the δ_i's as random variables, however, then the conditional null distribution of T gets well defined. Although the actual computation may become somewhat tedious, the conditional probability

$$\Pr(T \geq t \mid H_0, |x_1|, |x_2|, \ldots, |x_n|)$$

can be worked out. Thus, we can carry out one-sided or two-sided tail area tests in terms of the conditional null distribution of T.

The conditional test just described bears the distinctive hallmark of Sir Ronald. It was Fisher who amazed and mystified the statistical world with his sensational 2×2 conditional test of independence, and it was he who taught us how to set up a conditional test for the equality of two Poisson means.

In order to find the attained significance level of data vis à vis a null hypothesis H_0, we have to search for an appropriate test criterion T and then refer it to an appropriate sample space (the reference set) for determining the tail-area probability under the null hypothesis. In the present case the criterion is the sample total T and the reference set is the set of all samples of the type

$$(\pm|x_1|, \pm|x_2|, \ldots, \pm|x_n|) \ ,$$

where $|x_1|, |x_2|, \ldots, |x_n|$ are fixed at their observed values. Before we turn the searchlight of careful scrutiny on this mystifying conditional test, it will be useful to compare it with two familiar nonparametric tests of the null hypothesis $\mu = 0$. (See Kempthorne and Folks 1971, pp. 340–345.)

The sign test: Choose as the test criterion the number S of positive signs among $\delta_i = \operatorname{sgn} x_i$, $i = 1, 2, \ldots, n$. The null distribution of S is bin $(n, \frac{1}{2})$. One-sided or two-sided tests can then be made in terms of S.

The Wilcoxon signed-rank test: Instead of the sample total $T = \sum |x_1|\delta_i$, choose the statistic $W = \sum r_i\delta_i$ as the test criterion, where r_i is the rank of $|x_i|$ among $|x_1|, |x_2|, \ldots, |x_n|$. Observe that the range of variation of W is the set of alternate integers in the interval $[-n(n+1)/2,\ n(n+1)/2]$. Under the null hypothesis H_0, the two vectors $(r_1, r_2, \ldots, r_n)$ and $(\delta_1, \delta_2, \ldots, \delta_n)$ are stochastically independent and the δ_i's are iid ± 1 variables with equal probabilities. The conditional distribution of W, given $(r_1, r_2, \ldots, r_n)$, can, therefore, be easily worked out under H_0. Since $(r_1, r_2, \ldots, r_n)$ is always a permutation of $(1, 2, \ldots, n)$, it is clear that the null distribution of W is the same for all possible realizations of $(r_1, r_2, \ldots, r_n)$; in other words, the Wilcoxon statistic W is stochastically independent of the rank vector if the null hypothesis is true. Thus, the Wilcoxon test is not a conditional test in the sense the Fisher randomization test is. The sign test and the Wilcoxon test are typical examples of nonparametric, distribution-free, marginal tests.

Commenting on the three tests, Kempthorne and Folks (1971, p. 344) wrote: "Since the sign test uses only the signs of x_i, the Wilcoxon test uses only the signs and the ranks of $|x_i|$, and the Fisher test uses the x_i without condensation, the Fisher test is superior as a significance test." Thus, it seems that Kempthorne and Folks are giving poorer ratings to the sign test and the Wilcoxon test on the score that they violate the insufficiency principle to a greater extent than does the Fisher test. But how to measure the extent of such violations! Do any of these tests violate the sufficiency principle? Let us examine the question.

In the context of our nonparametric statistical model, the set of order statistics $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ is minimal sufficient. Each of the three test criteria T, S, and W can be written as a function of the order statistics, for example

$$T = \sum x_{(i)} = \sum |x_{(i)}|\delta_{(i)} \ ,$$
$$S = \tfrac{1}{2}(\sum \delta_{(i)} + n) \ ,$$
$$W = \sum r_{(i)}\delta_{(i)} \ ,$$

where $\delta_{(i)} = \operatorname{sgn} x_{(i)}$ and $r_{(i)}$ is the rank of $|x_{(i)}|$ among $|x_{(1)}|, |x_{(2)}|, \ldots, |x_{(n)}|$. Since the sign and the Wilcoxon tests are based on the marginal distributions of S and W, respectively (and, of course, on their observed values), there is no violation of the sufficiency principle in these cases—only the insufficiency principle is at stake.

In the case of the Fisher test, it may appear on the surface that the sufficiency principle has been violated in view of the fact that the conditioning statistic $(|x_1|, |x_2|, \ldots, |x_n|)$ is not a function of the minimal sufficient statistic $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$. If we carefully examine the conditional distribution of T given $(|x_1|, |x_2|, \ldots, |x_n|)$, then it will be clear that the conditional distribution depends on the sample $(x_1, x_2, \ldots, x_n)$ only through the order statistics. The sufficiency principle is sometimes interpreted as a requirement that the data ought to be first reduced to the minimal sufficient statistic (thus sieving out all the postrandomization impurities) and then the reduced data interpreted in terms of the marginal distribution of the minimal sufficient statistic. To satisfy the statistical intuition of such a purist we have only to point out that the Fisher test will remain unaltered if the conditioning statistic is chosen to be the ordered rearrangement of $|x_{(1)}|, |x_{(2)}|, \ldots, |x_{(n)}|$. The Fisher test does not violate the sufficiency principle.

The choice of the test criterion $T = \sum x_i$ and the choice of the conditioning statistic $(|x_1|, |x_2|, \ldots, |x_n|)$ are arbitrary elements in the Fisher test. For instance, we may want to condition T with respect to the statistic $(|x_1 + x_2 + \ldots + x_k|, |x_{k+1}|, \ldots, |x_n|)$ for some chosen k $(1 \leq k \leq n)$. It is easily seen that any such conditioning will make the null distribution of T dis-

tribution free. For instance, if $k = n$, then the conditional null distribution of T, given $|x_1 + \ldots + x_n| = d$, is uniform over the two-point set $\{-d, d\}$. With such a conditioning the one-sided test of the null hypothesis will result in a significance level of $\frac{1}{2}$ whenever the observed value of T is positive! Suppose we somehow convince ourselves that the only reasonable choice of a conditioning statistic is the one that corresponds to $k = 1$, the Fisher choice. (If $1 < k < n$, then the test procedure violates the sufficiency principle. The case $k = n$ is too ridiculous to deserve any serious consideration. And so on for any other conditioning statistic that one can think of.) Even then the question about the choice of the test criterion remains. Instead of $T = n\bar{x}$, why not choose the sample median $\tilde{x}$ as the test criterion? With the Fisher conditioning with respect to $(|x_1|, |x_2|, \ldots, |x_n|)$, the null distribution of $\tilde{x}$ is also distribution free. In our nonparametric setup, the sample median $\tilde{x}$ seems to be as reasonable a choice (as a test criterion) as the sample mean. Now, let us try to evaluate the significance level attained by the sample

$$(x_1, x_2, x_3, x_4, x_5) = (4, 7, 2, 3, 1).$$

The Fisher reference set consists of the 32 points

$$(\pm 4, \pm 7, \pm 2, \pm 3, \pm 1).$$

The observed sample mean is 3.4 and the median is 3. In the reference set there is only one point (viz., the sample itself) whose mean is at least as high as 3.4. In the same reference set, however, there are four points, namely, $(4, 7, \pm 2, 3, \pm 1)$ with median as high as 3. Therefore, with $\bar{x}$ as the test criterion the significance level (SL) of the data will be evaluated as 1/32, whereas with $\tilde{x}$ as the test criterion the data will be deemed to have attained SL = $\frac{1}{8}$. Note that every sample of five positive observations, irrespective of how far out or how scattered they are on the positive half-line, will be judged as significant (SL = 1/32) if $\bar{x}$ is the test criterion and not significant (SL = 1/8) if $\tilde{x}$ is the test criterion. With a sample of seven positive observations the SL will be 1/128 or 1/16 depending on whether $\bar{x}$ or $\tilde{x}$ is chosen as the test criterion. Consider the two samples $(-5, -4, -1, 6, 7, 8, 9)$ and (62, 63, 64, 65, 66, 67, 68). Does it make any sense to say that, with respect to the null hypothesis $\mu = 0$ with one-sided alternatives, the two samples are equally significant with SL = 1/16? But that is exactly what the Fisher test will do if $\tilde{x}$ is chosen to be the test criterion.

Let us take a short break from this ruthless cross-examination of the Fisher test with some speculation on Sir Ronald's later thoughts on the subject. The cross-examination will continue in Section 6.

5. DID FISHER CHANGE HIS MIND?

In all fairness to Sir Ronald we have to admit that, apart from making a passing reference to the randomization test method in Chapter III of *DE*, Fisher did not have much else to do with this kind of test procedure. Twenty-one years later, when Fisher came out with his last testament on statistics—*Statistical Methods and Scientific Inference* (*SI*)—he had apparently forgotten all about his randomization test method. In the winter of 1954–55, Fisher visited the Indian Statistical Institute for a couple of months and gave an extensive series of lectures based on the manuscript of *SI*. Those lectures profoundly influenced my own thinking on statistics. In *SI*, Fisher discussed the logic of inductive inference, his new outlook on significance testing, fiducial inference methods, likelihood methods of inference, and conditioning and recovery of ancillary information, but nowhere do we find any mention of randomization analysis of data. Randomization as an ingredient of statistical designs was mentioned only once, and that appeared in the following passage (Fisher 1956, p. 98):

> . . . whereas in the Theory of Games a deliberately randomized decision (1934) may often be useful to give an unpredictable element to the strategy of play; and whereas planned randomization (1935–53) is widely recognized as essential in the selection and allocation of experimental material, it has no useful part to play in the formation of opinion, and consequently in the tests of significance designed to aid the formation of opinion in Natural Sciences. [Note: The year 1934 refers to a Fisher article on randomization in card play and 1935–53 refers to *DE*.]

On the suggestion of an associate editor of *JASA* this passage is quoted in full so that the readers of the article can make up their own minds on the following.

Questions: Isn't it surprising that Fisher had no more to say about randomization in 1956? Was Fisher disassociating himself from the randomization test by not mentioning the method in *SI*? Does the remark about "formation of opinion" refer to postrandomization only?

It should be recognized that Fisher's views on significance testing underwent a major change during the period 1935–1956. On p. 77 of *SI* he made a clear distinction between tests of significance as used in natural sciences with tests for acceptance as in quality-control theory. According to him the dissimilarities (between the two methods) lie in the population, or reference set, available for making statements of probability. Let us quote Fisher (*SI*, p. 77) on this point:

> Confusion under this head has on several occasions led to erroneous numerical values; for, where acceptance procedures are appropriate the population of lots of one or more items, which could be chosen for examination, is unequivocally defined. The source of supply has an objective empirical reality. Whereas, the only populations that can be referred to in a test of significance have no objective reality, being exclusively the product of the statistician's imagination . . . . The demand was first made, I believe, in connection with Behrens' test of . . . significance . . . , that the level of significance should be determined by repeated sampling from the same population, evidently with no clear realization that the population in question is hypothetical, that it could be defined in many ways . . .; or, that an understanding, of what the information is which the test is to supply, is needed before an appropriate population, *if indeed we must express ourselves in this way*, can be specified. (Italics ours)

Again, on p. 91 of *SI*, Fisher quoted himself (from a 1945 *Sankhyā* article dealing with the fiducial argument) as follows:

> In recent times one often repeated exposition of the tests of significance, . . . , seems liable to lead mathematical readers astray, through laying down axiomatically, what is not agreed or generally true, that the level of significance must be equal to the frequency with which the hypothesis is rejected in repeated sampling of any fixed population allowed by hypothesis. This intrusive axiom, which is foreign to the reasoning on which tests of significance were in fact based seems to be a real bar to progress. . . .

It seems clear to me that, in 1956, Fisher's views on significance testing were somewhat close to the Bayesian position that the evidential content of data cannot be judged in sample space terms. Indeed, the 1945 quotation from Fisher might very well have been written by De Finetti himself. I am, therefore, not surprised at all that in *SI* Fisher mentioned neither the randomization test nor the lady-tasting-tea–type data analysis. For these are very extreme types of nonparametric data analysis in which the evidential meaning of the data is sought to be evaluated by referring it to a sample space that is formed by the statistician in his or her mind by imagining all the possible outcomes of the planned randomization input of the experiment. This will be made clearer in the next section.

Many of our contemporary statisticians are unaware of the fact that in the seventh edition of *DE* (1960), Fisher added what looks like a disclaimer in the form of a short section (Sec. 21.1; "Nonparametric" Tests) at the end of Chapter III. We quote this section in full.

> In recent years, tests using the physical act of randomization to supply (on the Null Hypothesis) a frequency distribution, have been largely advocated under the name of Nonparametric tests. Somewhat extravagant claims have often been made on their behalf. The example of this section, published in 1935, was by many years the first of its class. The reader will realize that it was in no sense put forward to supersede the common and expeditious tests based on the Gaussian theory of errors. The utility of such nonparametric tests consists in their being able to supply confirmation whenever, rightly or, more often, wrongly it is suspected that the simpler tests have been appreciably injured by departures from normality.
>
> They assume less knowledge, or more ignorance, of the experimental material than do the standard tests, and this has been an attraction to some mathematicians who often discuss experimentation without personal knowledge of the material. In inductive logic, however, an erroneous assumption of ignorance is not innocuous; it often leads to manifest absurdities. Experimenters should remember that they and their colleagues usually know more about the kind of material they are dealing with than do authors of textbooks written without such personal experience, and that a more complex, or less intelligible, test is not likely to serve their purpose better, in any sense, than those of proven value in their own subject.

Note Fisher's use of the phrase "physical act of randomization." The same phrase appears in the Fisher quotation in the opening paragraph of the previous section. Where is the physical act of randomization in the Fisher randomization test? The random entities $\delta_1, \delta_2, \ldots, \delta_n$ can hardly be called randomization variables. It is only under the null hypothesis that the δ_i's can be regarded as iid uniform ± 1 variables. The nonnull distribution of the δ_i's depends on the parameter of interest μ in a rather complex fashion. We should recognize the fact that in Section 21 of *DE* (1935) Fisher was not really concerned with the particular test situation that we have discussed in the previous section. He was talking about the problem of comparing two treatment effects under a wider hypothesis and was suggesting a (nonparametric) randomization analysis of data generated by paired comparisons on the basis of a physical act of randomization. In the next section we discuss this matter in some detail.

6. RANDOMIZATION AND PAIRED COMPARISONS

A scientist wants to test whether a so-called improved diet (treatment) is in effect superior to the standard diet (control). The scientist has 30 animals (subjects) with which to experiment. The scientist carefully pairs (blocks) the subjects into 15 homogeneous pairs. Let $\{(s_{1i}, s_{2i}): i = 1, 2, \ldots, 15\}$ be the set of 15 subject pairs. The subjects in each pair are of the same sex, come from the same litter, and so on. From each pair the scientist selects one subject for the treatment and the other one for control. The 30 responses (weight gain in so many weeks) are laid out as $\{(t_i, c_i): i = 1, 2, \ldots, 15\}$, where t_i and c_i are the responses of the treated subject and the control subject, respectively.

The scientist observes that

$$T = \sum t_i - \sum c_i$$

is a large positive number and also notes that

$$d_i = t_i - c_i > 0 \text{ for all } i \ .$$

The scientist, therefore, concludes that he or she has obtained very strong evidence in favor of the hypothesis H_1 that the improved diet is really superior to the standard diet. For measuring the strength of the evidence the scientist consults a statistician.

The statistician decides to make a one-sided test of significance of the null hypothesis H_0 (that the two diets are the same in their short-term weight-gain effects) on the basis of the scientist's data. The statistician also thinks that $T = \sum d_i$ is an appropriate test criterion in this case. For finding the significance level of the observed value of T, the statistician has to find the null distribution of T. So what the statistician needs now is a nice reference set.

The response difference d_i between the ith pair of subjects can be explained in terms of a possible treatment difference and other possible nuisance factors like subject differences (which the scientist tried his or her best to control by blocking), virus infection, loss of appetite, and many such uncontrollable factors that may have acted differently on the two subjects in the ith pair. If hypothesis H_0 is true, then there is no treatment difference; so the response difference d_i must be presumed to be caused by the previously mentioned nuisance factors. As Fisher explained in *DE*, randomiza-

tion enables the statistician to eliminate all these nuisance factors from the statistical argument. Let us see how this elimination is achieved in the present case.

Suppose the scientist had made 15 independent random decisions (on the basis of 15 tosses of a fair coin) as to which subject in the ith pair gets the improved diet ($i = 1, 2, \ldots, 15$). Having recorded the 15 response differences $d_1, d_2, \ldots, d_{15}$ and having computed $T = \sum d_i$, the scientist can speculate about a hypothetical rerun of the experiment in which all but one of the experimental factors (controllable or uncontrollable) are supposedly held fixed at the level of the last experiment—the same 30 animals exactly as they were at the commencement of the last trial, paired the same way into 15 blocks, exactly the same set of animals coming down with the same kind of virus infections with the same effects on them, and so forth. The only thing that is allowed free play in the hypothetical rerun of the experiment is the random allocation of treatment—the fair coin has to be tossed again 15 times. If H_0 is true, then the response difference $d_i = t_i - c_i$ for the ith pair must have been caused by the nuisance factors (subject differences, virus infection, etc). In the hypothetical rerun of the experiment all such nuisance factors are supposedly held fixed at the past level. Therefore, in the new experiment the response difference for the ith pair can take only two values d_i or $-d_i$, depending on whether the treatment allocation for the ith pair is the same as in the past experiment or is different. If we denote the response differences for the hypothetical experiment by $(d'_1, d'_2, \ldots, d'_{15})$ then it is clear that, under the null hypothesis H_0, the sample space (for the response differences) is the set R of 2^{15} points (vectors)

$$R = \{(\pm d_1, \pm d_2, \ldots, \pm d_{15})\} ,$$

with all the points equally probable. This is the reference set that the statistician was looking for. Let $T' = \sum d'_i$. The significance level of the data

$$\text{SL} = \Pr(T' \geq T | H_0)$$

is now computed as follows.

The statistician looks back on the data and notes that $d_i > 0$ for all i. Therefore, $T' \geq T$ if and only if $d'_i = d_i$ for all i. Hence, $\text{SL} = \frac{1}{2}^{15}$. This is randomization analysis of data in its classical form.

The rest of this section is devoted to an evaluation of this particular data analysis. The evaluation is laid out in the form of a hypothetical sequence of remarks and counterremarks by the statistician, the scientist, and the author.

Statistician: Observe that the randomization test argument does not depend on any probabilistic assumptions. The randomization probabilities are fully understood and are completely under control. I do not have to assume that the treatment-allocation process was like a sequence of 15 Bernoulli trials with $p = \frac{1}{2}$. Surely, I can regard that as demonstrably correct. In this argument there is no mention of a population. The experimental animals do not have to be regarded as a random sample from a population of animals. This test is an ultimate nonparametric test. Not only do we not have to deal with model parameters, we do not have to contend with even a statistical model. There is no mention of a sample space X equipped with a σ-field A of events and a family P of probability measures, no measurement errors, no mention of a sequence of iid random variables with an unknown distribution function.

Scientist: I am greatly puzzled by your data analysis. Your analysis seems to depend only on the randomization probabilities and the observed fact that $d_i > 0$ for all i. The fact that the test criterion $T = \sum d_i$ attained a rather large value in this case does not seem to enter into the probability evaluation of $\frac{1}{2}^{15}$.

Author: Suppose we choose the median of $d_1, d_2, \ldots, d_{15}$ to be the test criterion instead of $T = \sum d_i$. The significance level of the data will then be evaluated as $\frac{1}{2}^8$. How can we explain the big difference between $\frac{1}{2}^{15}$ and $\frac{1}{2}^8$?

Scientist: I do not understand the relevance of the randomization probabilities. Why is it so crucial that the coin with which I made the treatment allocation be a fair coin? Suppose I had used a biased coin with $p = \frac{1}{4}$. Suppose for the ith pair (s_{1i}, s_{2i}) of experimental animals my treatment allocation was (t, c) or (c, t) depending on whether the ith toss of the biased coin resulted in a head or a tail. How significant would my present data have been then?

Author: Let me answer the question. The hypothetical rerun of the experiment will be defined as before, but this time the biased coin will define the randomization scheme. The reference set for $(d'_1, d'_2, \ldots, d'_{15})$ will still be the same set R. Note that in this case $\Pr(d'_i = d_i) = \frac{1}{4}$ or $\frac{3}{4}$ depending on whether, in the original experiment, the response difference d_i was associated with the (t, c) or the (c, t) treatment allocation. Therefore, the level of significance will be evaluated as

$$\text{SL} = \Pr(T' \geq T | H_0) = (\tfrac{1}{4})^m (\tfrac{3}{4})^{15-m} ,$$

where m is the number of (t, c) allocations in the original experiment. The larger the value of m is, the more significant are the data.

Scientist: This is patently absurd. How can the SL depend so largely on such an irrelevant data characteristic as m? It is relevant to know that the 30 animals have been paired into 15 homogeneous blocks. The manner of my labeling the two animals in the ith block as (s_{1i}, s_{2i}) does not seem to be of much relevance. The number m of treatment allocations of the type (t, c) seems to be of no consequence at all. I have not been asked about all the background information that I have on the problem. For instance, I happen to have made a nutrition analysis of the two diets. I know that the improved diet has a much higher protein content and is very rich in vitamins C and D. I know the results

of several past experiments on the same set of animals when they were fed the standard diet. I know that six animals came down with virus infections during the experiment and that five of them were fed the improved diet. I am amazed to find that a statistical analysis of my data can be made without reference to these relevant bits of information.

Statistician: You are trying to make a joke out of an excellent statistical method of proven value, a method that originated in the mind of one of the two (Fisher and Einstein) really outstanding men of genius that the world has seen in this century. Your criticisms are based on an extreme example and then on a misunderstanding of the very nature of tests of significance. Tests of significance do not lead to probabilities of hypotheses. I do not believe in "belief probabilities." I do not believe that any useful purpose can be served by trying to quantify your knowledge in the form of a belief probability. Go to a Bayesian if you wish to make any input of your subjective beliefs in the data analysis process. It does not make much sense to set up a statistical model for the purpose of analyzing experimental data. The randomization analysis of data is so simple, so free of unnecessary assumptions that I fail to understand how anyone can raise any objection against the method. In the case of the present experiment you have in effect tossed a fair coin 15 times, have you not? So why confuse the issue by bringing in the case of an absurdly biased coin with $p = \frac{1}{4}$? Note that the probability of $\frac{1}{2}^{15}$ that I have computed for you is a gambler's probability, a frequency probability, a propensity measure of a well-defined physical system. A belief probability it is not.

Scientist: Your probability of $\frac{1}{2}^{15}$ is defined in terms of a hypothetical experiment, a rerun of the original experiment with everything (repeat everything) but the randomization part fixed at the level of the original experiment. But how can you even think of such an utterly impossible experiment? My experimental animals have changed—one of them died last week—the weather has changed, the virus epidemic is gone. I do not see how you can claim any objective reality for the randomization probability of $\frac{1}{2}^{15}$. In any case, I knew all along that the null hypothesis could not possibly be true. So any probability computed under the supposition that the null hypothesis is true cannot have much of an objective reality.

Author: The computation SL = $\frac{1}{2}^{15}$ was based on the supposition that in the hypothetical rerun of the experiment all the 2^{15} treatment-allocation patterns are equally probable. It is not clear from the argument that the scientist had to make all the 2^{15} possible allocations equally probable in the original experiment.

Scientist: This is a good time for me to confess that in fact I did not randomize over the full set of 2^{15} possible allocations. As a scientist I have been trained to put as much control into the experimental setup as I am capable of, to balance out the nuisance factors as far as possible. After carefully blocking the 30 subjects into 15 nearly homogeneous pairs, I could still detect differences within the subject pairs. There were differences in weight, height, some relevant blood characteristics, and a few other relevant features. I wanted the set of 15 treated subjects to be nearly equal to the set of 15 control subjects in some group characteristics like average weight, average height, and so on. I worked very hard on the project of striking a perfect balance between the treatment and the control groups. Finally, I found two such complementary groups and then decided on the treatment/control allocation to the two groups by a mental process that may be likened to the toss of a fair coin. I wonder what the significance level of my data is going to be in the light of this confession.

Statistician: Had I known about this before, I would not have touched your data with a long pole. Now the reference set for $(d'_1, d'_2, \ldots, d'_{15})$ consists of only the two points

$$(d_1, d_2, \ldots, d_{15}) \quad \text{and} \quad (-d_1, -d_2, \ldots, -d_{15}) ,$$

and the significance level

$$\text{SL} = \Pr(T' \geq T | H_0)$$

works out to be $\frac{1}{2}$ if $T > 0$ and 1 if $T \leq 0$. Your data is not significant at all.

Scientist (utterly flabbergasted): But my experiment was better planned than a fully randomized experiment, was it not? With my group control (in addition to the usual local control) I made it much harder for $T = \sum t_i - \sum c_i$ to be large in the absence of any treatment difference. In spite of this careful global control, I found that T is a large positive number and that every $t_i > c_i$. And you are telling me that, under the null hypothesis, it is as easy to get a result as significant as mine as it is to get a head from a single toss of a symmetric coin!

Statistician: My good man, you must realize that your experiment is no good. The prerandomization that you had carried out was not wide enough; the randomization sample space has only two points in it with a uniform probability distribution under the null hypothesis. Thus, the only attainable significance levels are $\frac{1}{2}$ and 1. Your experiment is not informative enough. I wish you had consulted me before planning your experiment. It appears that you do not have a clear understanding of the role of randomization in statistical experiments.

7. CONCLUDING REMARKS

So the randomization argument foundered on the rocks of restricted and unequal probability randomization. The statistician had the last word but lost the argument. The statistician was clearly wrong in characterizing the scientist's one-toss randomized experiment as uninformative. During the last 15 years, I have heard three very eminent statisticians characterizing the one-

toss experiment as uninformative on the score that the sample space has only two points in it. That this cannot be so is easily seen as follows.

An urn contains two balls that are either both white or both black. The draw of a single ball from the urn is then fully informative, although the sample space has only two points in it. If this example seems to be too artificial, then consider the case of an urn in which the proportion of white balls is either $\frac{1}{4}$ or $\frac{3}{4}$. Consider the sequential sampling plan that requires drawing of balls one at a time and with replacements until the likelihood ratio either exceeds 100 or falls below 1/100. Suppose the outcome of this experiment is recorded as "below 1/100" or "above 100." This is a highly informative experiment with only two sample points in it.

It should be noted that the sample space of the experiment performed by the scientist had a huge number of points in it. The statistician took a thin cross-section of the sample space (after holding fixed all the relevant factors like subjects, treatment effects, recognizable nuisance factors, and error terms) and then found only two points in it. No wonder the scientist failed to understand the argument.

The scientist was correct in questioning the relevance of randomization at the data analysis stage. Prerandomization injects an element of uncertainty about the actual experimental layout. But that uncertainty is removed once the scientist goes through the randomization ritual early in the game. At the data analysis stage, why is it still necessary to find out about the details of the actual randomization process? The randomization exercise cannot generate any information on its own. The outcome of the exercise is an ancillary statistic. Fisher advised us to hold the ancillary statistic fixed, did he not?

Our statistician is a most ardent admirer of R.A. Fisher. But he does not like the postfiducial (1936–62) Fisher. During the last 27 years of his astonishing career, we find Sir Ronald entertaining such counterrevolutionary thoughts as the conditionality and the likelihood principle and toying with the half-baked Bayesian idea of fiducial probability distribution.

We have noted earlier how the sufficiency principle rejects postrandomization analysis of data. Similarly, the conditionality principle (see Basu 1975 for more on this) rejects prerandomization analysis of data. In view of Fisher's postfiducial rethinking on statistical inference, it was almost inevitable for him finally to insert that astonishing short section on nonparametric tests in the seventh edition of *The Design of Experiments.*

[*Received March 1979. Revised October 1979.*]

REFERENCES

Basu, D. (1975), "Statistical Information and Likelihood" (with discussions), *Sankhyā*, Ser. A, 37, 1–71.
——— (1977), "On the Elimination of Nuisance Parameters," *Journal of the American Statistical Association*, 72, 355–366.
——— (1978), "On the Relevance of Randomization in Data Analysis" (with discussion), in *Survey Sampling and Measurement*, ed. N.K. Namboodiri, New York: Academic Press, 267–339.
Fisher, R.A. (1956), *Statistical Methods and Scientific Inference*, Edinburgh: Oliver and Boyd.
——— (1960), *The Design of Experiments* (7th ed.), Edinburgh: Oliver and Boyd.
Hodges, J.L., Jr., and Lehmann, E.L. (1973), "Wilcoxon and *t*-Test for Matched Pairs of Typed Subjects," *Journal of the American Statistical Association*, 68, 151–158.
Kempthorne, O. (1952), *The Design and Analysis of Experiments*, New York: John Wiley & Sons.
——— (1955), "The Randomization Theory of Experimental Inference," *Journal of the American Statistical Association*, 50, 946–967.
——— (1966), "Some Aspects of Experimental Inference," *Journal of the American Statistical Association*, 61, 11–34.
——— (1974), "Sampling Inference, Experimental Inference and Observations Inference," Paper presented at the Mahalanobis Memorial Symposium on Recent Trends of Research in Statistics, Calcutta, India.
——— (1975), "Inference for Experiments and Randomization," in *A Survey of Statistical Designs and Linear Models*, ed. J.N. Srivastava, Amsterdam: North Holland Publishing Co., 303–331.
——— (1977), "Why Randomize?" *Journal of Statistical Planning and Inference*, 1, 1–25.
Kempthorne, O., and Folks, J.L. (1971), *Probability Statistics and Data Analysis*, Ames: Iowa State University Press.
Pitman, E.J.G. (1937), "Significance Tests Which Can Be Applied to Samples From Any Population III. The Analysis of Variance Test," *Biometrika*, 29, 322–335.
Wald, A. (1950), *Statistical Decision Functions*, New York: John Wiley & Sons.
Warner, S.L. (1965), "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," *Journal of the American Statistical Association*, 60, 66–69.

Comment

DAVID V. HINKLEY*

Basu has provided us with an interesting and provocative critique of significance tests related to randomized experiments. It does seem to be true that there is not a unified Fisherian mathematical theory of significance tests. This should not be surprising, however, since Fisher was wont to warn of the dangers of

* David V. Hinkley is Professor and Chairman, Department of Applied Statistics, and Professor, Department of Theoretical Statistics, University of Minnesota, Minneapolis, MN 55455.

September 1980, Volume 75, Number 371
Theory and Methods Section

routine, formal application of mathematical statistics without very careful regard for scientific context and operational meaning. Indeed, one might view Basu's paper as an illustration of Fisher's warning.

In terms of Fisher and randomization, the first four sections of the paper require little comment, since they deal with the separate topic of permutation and rank tests. Nevertheless, it is important to point out a fallacy in Basu's criticism of nonunique significance level (SL) in Section 4: The data as such do not possess an SL, which instead attaches to a particular statistic. Moreover, it is important to recognize that in Section 4 there is only an abstract null probability model—not a general model—so that the statistician has no basis for choice of statistic: The scientist must specify the relevant statistical measure. The role of the statistician here is to ensure that a valid, operational interpretation of the chosen statistic can be, and is, made.

After confessing to a "ruthless cross-examination" of the wrong topic—the non-Fisherian nonparametric tests of Section 4—Basu suggests that Fisher's silence in 1956 may be used to condemn the randomization test. This speculation seems unwarranted on two counts. First, I do not think that Fisher ever did recommend the randomization test for analysis of data, but rather that he introduced it as a device for demonstrating that randomization validates the usual normal-theory methods of analysis. This notion seems clear in Yates (1933), for example. Unfortunately, Basu has not chosen to discuss the connection between randomization and the validity of mathematical models. The second point is that Fisher's views on randomization would have been so widely known and accepted after 25 years that it would not have seemed necessary to repeat them in a book on statistical inference for parametric probability models. Fisher did repeatedly assert that randomization could guarantee the relevance of such abstract models—including the seemingly innocuous model in Section 4—but he realized that randomization was "sufficient" (CP 204)[1] rather than necessary, since often "Nature has done the randomization for us" (CP 212). These last remarks should be borne in mind when reading the amusing developments in Section 6.

What are we to make of the Statistician and the Scientist? They are certainly an entertaining addition to the literature, but hardly enlightening or enlightened. The first serious issue seems to be that of the biased-coin design, for which the author provides the SL. Surely the SL given is an appropriately cautious evaluation in the worst case in which the experimenter knowingly takes advantage of the bias—cheats, that is. But apparently the Scientist did not cheat ("my labeling . . . does not seem to be of much relevance"), so that in effect the treatments were allocated at random within each pair: Nature has done the randomization for us. Thus the usual analysis is presumably valid, and if a randomization SL were computed it would still be $\frac{1}{2}^{15}$. The Statistician has apparently mistaken Fisher's "sufficient" for "necessary."

What follows after the biased-coin episode is a series of irrational remarks and misunderstandings. The Statistician's dogmatic attitude is hardly characteristic of the statistician who inspired the Rothamsted song "Why! Fisher can always allow for it" (Box 1978, pp. 138–139).

What was Fisher's position on randomization and the induced distribution of statistics? While this is not entirely clear down to the last detail, I think it clear enough to suggest that Basu has missed the point. For a brief introduction to the relevant parts of Fisher's work, see the lectures by Holschuh, Picard, and Wallace in Fienberg and Hinkley (1980). Highly informative and balanced accounts of the issues may be found in Yates (1970), particularly the 1965 Berkeley Symposium paper reviewing experimental design. As I see it, the purpose of randomization in the design of agricultural field experiments was to help ensure the validity of normal-theory analysis. Nature was not in the habit of doing the randomization. Studies by Tedin (1931) and others on uniformity trial data showed that for systematic designs (such as that finally described by Basu's Scientist) the usual properties of t and F tests did not hold in an operational sense. Thus standard significance tests were invalidated. "Student," among others, correctly pointed out that effects could be more precisely estimated from carefully chosen systematic designs. But, said Fisher, this was of no use if the estimated precision were too high, higher even than the valid estimates obtained from randomized experiments. Thus, for some systematic designs, the computed normal-theory SL corresponding to a theoretically precise effect was in fact appreciably larger than the "real" SL. This is exactly what could happen in the case of the two-point design of Basu's Scientist, although it probably did not if nature has randomized. With the limited information given us by Basu, we cannot give a reliable standard error for the Scientist's accurate estimate $\bar{d}$, at least not one with a clear operational meaning. The apparently silly SL values ($\frac{1}{2}$ and 1) are a warning of possible difficulty, surely, nothing more.

The empirical evidence confronting Fisher certainly suggested the necessity of randomization in most field experiments, if the standard methods of analysis were to be used. In recent years it has become apparent that relatively simple spatial models can often account for some of the effects that randomization was designed to overcome; see Bartlett (1978), for example. Complete or partial failure to randomize can have adverse effects in other areas too, for example, in survey sampling in which systematic grids are randomly positioned on a sampling frame. In such a case systematic effects can accidentally (or purposely) change the variation, as I have seen myself. Cochran (1977, Ch. 8) discussed

[1] Fisher's papers are referred to by their numbering in the *Collected Papers* (Fisher 1974).

this problem in detail. For informative accounts of the importance of randomization in medical and public-policy studies, respectively, see Chapters 9, 10, and 11 of Bunker, Barnes, and Mosteller (1977) and Gilbert, Light, and Mosteller (1977). In all areas, the randomization distribution literally induced by the experimental randomization is of value in assessing the validity of a standard analysis. This, I think, is Fisher's message.

The final substantial issue of Basu's paper is that of the ancillarity of the design outcome. Technically Basu is quite correct, if the randomization has validated a parametric model—the design outcome is then ancillary *by design!* It would, however, be as well not to forget the purpose of an ancillary statistic, since otherwise we are merely playing with abstract mathematical definitions. An ancillary statistic indicates the set of comparable cases against which to judge the observed sample and the statistical summary thereof. Usually "comparable cases" is taken to mean "equally informative samples" in some appropriate sense, as in Fisher's brief comments on the 2×2 table (CP 205). Admittedly this is not mathematically precise, but it seems to have the merit of common sense. It is often unnecessary, and sometimes plain foolish, to take an infinitesimal slice through an abstract space as the set of comparable cases, although the mathematical definition of ancillarity would require this. Lest Basu think that Fisher has been caught with his conditional pants down here, let me suggest that Fisher implicity invoked conditionality in his criticism of Knut-Vik Squares, which in Tedin's (1931) analyses correspond to an ancillary set of real import.

For a more constructive use of design ancillarity, consider a randomized block design (RBD) with four replicates of four treatments. Suppose that in the particular physical layout the selected design coincides with a 4×4 Latin Square and that the accidental block structure corresponds to a noticeable effect not due to the treatments. Here my qualitative notion of ancillarity would suggest that we analyze the experimental data as coming from a Latin Square design, that is, treat the 4×4 Latin Squares as that subset of RBD's that constitute the set of comparable cases. The Latin Square analysis would be exactly equivalent to using a covariate-adjusted RBD analysis with accidental block totals as covariates. (There is of course a question as to whether randomization validates the latter analysis.)

This short discussion has necessarily focused on my major misgivings with Basu's interesting paper. As to whether randomization tests are *logically* viable, I think Basu has not made a case. There may be no case in logic if, with John Clerk Maxwell, we believe that "the true logic for this world is the calculus of Probabilities." What we need to know is: Which probabilities?

[*Received December 1979.*]

REFERENCES

Bartlett, M.S. (1978), "Nearest Neighbour Models in the Analysis of Field Experiments" (with discussion), *Journal Royal of the Statistical Society*, Ser. B, 40, 147–174.

Box, J.F. (1978), *R.A. Fisher. The Life of a Scientist*, New York: John Wiley & Sons.

Bunker, J.P., Barnes, B.A., and Mosteller, F. (1977), *Costs, Risks, and Benefits of Surgery*, New York: Oxford University Press.

Cochran, W.G. (1977), *Sampling Techniques* (3rd ed.), New York: John Wiley & Sons.

Fienberg, S.E., and Hinkley, D.V. (1980), *R.A. Fisher: An Appreciation, Lecture Notes in Statistics*, New York: Springer-Verlag.

Fisher, R.A. (1974), *Collected Papers of R.A. Fisher*, Adelaide, Australia: University of Adelaide. (Papers are referred to by number.)

Gilbert, J.P., Light, R.J., and Mosteller, F. (1977), "Assessing Social Innovations: An Empirical Base for Policy," in *Statistics and Public Policy*, ed. W.B. Fairley and F. Mosteller, Menlo Park, Calif.: Addison-Wesley.

Tedin, O. (1931), "The Influence of Systematic Plot Arrangements Upon the Estimate of Error in Field Experiments," *Journal of Agricultural Science, Cambridge*, 21, 191–208.

Yates, F. (1933), "The Formation of Latin Squares for Use in Field Experiments," *Empire Journal of Experimental Agriculture*, 1, 235–244. (Also reprinted in Yates 1970.)

——— (1970), *Experimental Design. Selected Papers of Frank Yates, C.B.E., F.R.S.*, London: Griffin.

Comment

OSCAR KEMPTHORNE*

Basu states that we have no satisfactory answer to the question, "Why randomize?" Various workers have attempted to give "satisfying" partial answers to this question. Surely, Fisher (7th ed., 1960) did so, with extensive exposition in his *The Design of Experiments*. Then we can examine various writings of the 1930's.

We can cite Greenberg (1951) with the question as the title, as also was the title of the cited Kempthorne

* Oscar Kempthorne is Professor of Statistics and Distinguished Professor in Sciences and Humanities, Iowa State University, Ames, IA 50011.

September 1980, Volume 75, Number 371
Theory and Methods Section

(1977). It would be useful to have an attempted exhaustive bibliography of the topic.

It is obvious that expositions that some regard as carrying *some* real force are not so regarded by Basu and, for example, by Harville (1975). Is there any possibility that discussion will resolve the disagreement? I believe not. But I do believe that discussion is useful. All of us surely subscribe to the absolute necessity of critical examination of our ideas.

My discussion consists of two parts: (a) reactions of Basu's essay and (b) a few comments on the nature and role of randomization.

Basu writes entertainingly, perhaps, but not informatively. He poses the question, "Can the Fisher randomization test pass the test of common sense?" We must, I suggest, force Basu to be explicit and clear. What is this "common sense" that Basu refers to? Presumably, it is Basu's "common sense." The philosophy of statistics is plagued with writers who talk about "the probability" only to tell us that they mean "my probability"; now we have "common sense," but it is "Basu's common sense."

Section 2 given us prerandomization, postrandomization, and unrecorded randomization. This discussion is irrelevant. But it is useful, perhaps, to make a remark. It is ludicrous that Basu, a keen bridge player, does not, it seems, give a role to postrandomization. Let Basu play poker with significant (to him) payoff. Then if he does not use some sort of postrandomization, maybe very informal, he will be "cleaned out." I speak from past personal experience of playing social, but nontrivial (it now bores me!) poker. I regret that I surmise that many of those who write about gambling do not practice it. Section 2 is a red herring.

Section 3 discusses the sufficiency principle. As he has written, and others too numerous to cite, this is a data-reduction principle. It was used by Fisher in his (inadequate) formulation of tests of significance and reached its summit for Fisher in his fiducial inversion (which I shall not discuss). Problems with both of these led to the also inadequate formulation of use of ancillaries. The Fisher prescription, "The (Basu-titled) Insufficiency Principle," reached its summit only with fiducial inference, which was, in fact, the only real inference that Fisher espoused. It is in these terms, I suggest, that the later Fisher must be examined.

Section 4 discusses the Fisher randomization test on the basis of the cited Kempthorne-Folks presentation. My only regret here is that Basu did not examine, it seems, an article by Kempthorne and Doerfler (1969). I found Basu's remarks on the test not violating the sufficiency principle interesting and possibly a justification of the quoted remark of Kempthorne and Folks about condensation. Also, it was comforting that Basu seems to conclude that $k = 1$ gives "the only reasonable choice of a conditioning statistic." On the choice of test criterion, the naive idea I had was that the alternative is a uniform shift so that the sample total contains, perhaps, the maximum total shift. Here, there surely is a question. To this I add that my own use of the randomization test is in the experimental setting in which the alternative hypothesis is that a treatment adds some quantity Δ to each and every unit to which it is assigned. Also on a technological level, a question of interest is whether the treatment gives a gain when applied to *all* the experimental units.

Section 5 discusses whether Fisher changed his mind. This has perplexed others, including me (Kempthorne 1966, 1974, 1975). I have commented (1975) on what appeared to me to be outright inconsistencies in Fisher's whole output. It serves no useful purpose to try to psychoanalyze this phenomenon, I suggest. I do suggest, however, that we frankly admit the occurrence of these inconsistencies. Clearly, Fisher (1956) was writing in part a polemic against acceptance procedures. In connection with one quotation, it is obvious that the population in a randomization test of a randomized experiment is "the product of the statistician's imagination." (Why Fisher should include "exclusively," I do not know.) *It is not clear* to me that in 1956 Fisher had the position that "evidential content of data cannot be judged in sample space terms." On the matter of the "lady tasting tea" and randomized experiments, Fisher (1956) is, I judge, entirely silent, and that is surely a mystery. I have to state my opinion that I do not find Basu's psychoanalysis clear or convincing. On the Fisher (1960) quotation, Fisher is merely polemic, for some unknown reasons. In fact one can use Fisher's words against Fisher. One does *not* have knowledge of distribution. If one did, one would not be involved in transformation search, for instance. Any supposed statistician who believes he or she knows the model, for example, of normality and independence, is not a real statistician; that is surely obvious. The only interpretation of this is that Fisher was polemicizing, against what we can guess, but with no profit.

With respect to Basu's writing on "the physical act of randomization," I believe Basu is merely *plain wrong*. In a randomized experiment, the δ_i's have a known distribution whether or not the null distribution holds.

Section 6 gives in the first part what is, or should be, routinely taught on the randomized pair trial. This has been known for decades and an elementary substantively oriented exposition is that of Kempthorne (1961).

In the second part Basu gives a hypothetical interchange of a statistician and a scientist and the author. I suggest that this serves *no* useful purpose. Comments on sentences of this interchange follow.

1. *Scientist:* "The fact that . . . $T = \sum d_i$ attained a large value . . . does not seem to enter."

Comment: Precisely! When is T large? With reference to what is T large? If one has external information on the possible magnitude of T, then one will have an idea of what values of T are large. If one is a Bayesian, one *claims to know* the possible distribution of T. Clearly,

if one is in this situation, one does not randomize. I believe Basu has no familiarity with the problems of evaluating drugs for human illnesses, of evaluating diets on humans or mice or whatever. He does not see the variability among humans "treated alike." Why, is a mystery to me. Basu seems to say: If you are an expert on cancer, then you know a probability model. Otherwise, you should withdraw from the field. I regret that I find the lack of knowledge that underlies Basu's thesis rather surprising, incongruous, and deplorable. I would like Basu to take up a "very small" branch of investigative science, learn all the available background, and then design and conduct, with aid, of course, his own research program. Because Basu is highly competent, I believe, in the game of bridge, I would like him to make a comparative trial of two bridge systems. How would he do this? He surely has as good a background as any bridge player or writer. It is the absence of any effort on a problem outside the WFFing (constructing well-formed-formulas) of mathematical statistics that concerns me. As I have said before, we must be skeptical of individuals who write books on cooking but have never made a meal in a kitchen.

2. *Scientist:* "Why is it crucial that the coin . . . be a fair coin?"

Comment: From Basu's viewpoint, this is obviously irrelevant. From the viewpoint of the investigator who is not a Bayesian, the situation is different. If one does not regard experimentation as a process of investigation, with the value of the process being determined, partly at least, by its operating characteristics, the question is irrelevant. But a scientist who does not care about the operating characteristics of his or her observation procedure is a "pretty poor" scientist. This is my opinion, of course, but one that is shared by the great bulk of scientists, I am totally sure. Indeed, the question with respect to any substantive experimental outcome is whether other scientists can duplicate the results. This is all very elementary, and I will not waste valuable journal space discussing it. The question is irrelevant to Basu, because one should not use any coin. But for the person who accepts the idea that operating characteristics of a procedure are important and who regards significance tests as evidential, the answer is obvious. If one uses a collection of plans, of which one plan has probability .99, then the significance level regardless of outcome and regardless of whether there is a real treatment difference will be equal or exceed .99 with probability .99. One then has to discard the idea of significance tests—at least as they are used at present. If there are M plans, then using these with equal probability gives the possibility with huge treatment effects of obtaining a significance level of $1/M$. So, for the significance tester, there is value to equal probabilities, or the "fair coin." From another point of view, the use of equal probabilities gives estimates with nice properties, an analysis of variance with nice properties, and, surely, a valid use of the central limit theorem for the distribution of the test criterion in comparative trials of reasonable size. The reply of the author does not consider, I think, operating characteristics.

3. *Scientist:* "This patently absurd . . . ."

Comment: I can say, equally as Basu, that his writings on randomization are "patently absurd"—but, of course, this does not lead to improved understanding. Basu credits the scientist with all sorts of "background information." The scientist "knows etc." The scientist and Basu are entitled to "be amazed to find that a statistical analysis of my data can be made without reference to these relevant bits of information." Why? Because if the scientist really has these "bits of information" a decent statistician will attempt to take them into account. *No one claims that the randomization test of significance is the beginning, middle, and end of statistical analysis.* Finally, I must ask the question, "Has Basu worked intimately with any scientists with a real problem (as opposed to a circus trainer with 10 elephants)?"

4. *Scientist:* "But how can you even think of such an utterly impossible experiment?"

Comment: I, Kempthorne, can! A very simple answer! I follow this with a question to Basu, related to the very interesting arcane mathematics he sometimes does. "How can you, Basu, even think of observing a real number exactly?" Each of us has a mental problem. Let us rest the matter there.

5. *Scientist:* ". . . I did not randomize over the full set . . . ."

Comment: For me as a randomizing significance tester, that presents no problems. Tell me what your randomization frame was, and I can proceed. I may well find that to interpret your results a repeated sampling principle is useless. I, or rather you, have to supply a prior and a probability distribution. This is, perhaps, no problem for you. But it is for me, because now I have to assess for myself how much belief to hold with respect to your opinion. That is the rub!

6. *Statistician:* "Your data are not significant at all."

Comment: Right on! Your data are not significant to me. They may be to you, of course.

7. *Basu:* "So the randomization argument founders on the rocks of restricted and unequal probability randomization."

Comment: I do not see the claimed foundering.

8. *Basu:* "The one toss experiment (is) uninformative . . . ."

Comment: It is uninformative to me in the absence of forcing relevant and supported prior or external information. With such information, the actual ex-

periment is only a part of the total information. What is there to argue about?

9. *Basu:* "The outcome of the (randomization) exercise is an ancillary statistic."

Comment: Yes and no. This outcome does not depend on the probability model, but if one does not know the probability model, one cannot (or should not) characterize the randomization outcome as ancillary. Furthermore, Basu's own work (not cited, but very well known) shows that there are huge difficulties in strict formulation of ancillary statistics.

I have one final comment about Fisher (1956). It is clear from Chapter III that Fisher envisaged various forms of inference from tests of significance to distributions on unknown parameters. He did *not*, then, reject tests of significance in 1956. Furthermore, there is *no* evidence that he rejected his "lady tasting tea" example.

I close with the statement (which will be unknown to most readers of this journal) that Basu and I are very deep friends. The argumentation and the comments I make in this article must be interpreted with that background.

[*Received December 1979.*]

REFERENCES

Fisher, R.A. (1956), *Statistical Methods and Scientific Inference*, Edinburgh: Oliver and Boyd.

——— (1960), *The Design of Experiments* (7th ed.), Edinburgh: Oliver and Boyd.

Greenberg, B.G. (1951), "Why Randomize?" *Biometrics*, 7, 309–322.

Harville, D.A. (1975), "Experimental Randomization: Who Needs It?" *The American Statistician*, 29, 27–31.

Kempthorne, O. (1961), "The Design and Analysis of Experiments With Some Reference to Educational Research," in *Research Designs and Analysis, Second Annual Phi Delta Kappa Symposium on Educational Research*, 97–126.

——— (1966), "Some Aspects of Experimental Inference," *Journal of the American Statistical Association*, 61, 11–34.

——— (1974), "Sampling Inference, Experimental Inference and Observations Inference," Paper presented at the Mahalanobis Memorial Symposium on Recent Trends of Research in Statistics, Calcutta, India.

——— (1975), "Inference for Experiments and Randomization," in *A Survey of Statistical Designs and Linear Models*, ed. J.N. Srivastava, Amsterdam: North-Holland Publishing Co., 303–331.

Kempthorne, O., and Doerfler, T.E. (1969), "The Behaviour of Some Significance Tests Under Experimental Randomization," *Biometrika*, 56, 231–247.

Comment

DAVID A. LANE*

The scientist's experimental results contain evidence bearing on the superiority of the improved diet. He asks the statistician to evaluate this evidence. The statistician answers by computing a significance probability, $\frac{1}{2}^{15}$, by means of Fisher's randomization test. The scientist is baffled:

How can the evidence in his results be measured by a computation that ignores so much relevant information: the magnitude of the difference in weight gain between the two groups of animals, the ingredients of the two diets, previous experience with the standard diet, knowledge of the experimental animals gathered before and during the experiment, the mechanisms of the growth process, and so forth?

The statistician's computation refers to a biologically irrelevant feature of the experiment, the physical properties of the device determining the assignment of animals to diet; how can such a computation connect to the biological problem of the superiority of the improved diet?

The answer to the first question is clear: The statistician's significance probability cannot summarize completely the evidence in the scientist's experiment. The scientist cannot get something for nothing. If the scientist wants to assess what his experimental results imply about the effects of his improved diet and the nature of the growth process, he must analyze them in terms of a statistical model that describes as much as possible of what he knows about the biology of the experiment. But there are rhetorical as well as inferential issues involved in discussing an experiment. One of the scientist's goals is to obtain public confirmation for the superiority of the improved diet. If this can be accomplished with a minimum of fuss and assumption, preliminary to the detailed, model-based analysis, and without contradicting explicitly or implicitly the results of that analysis, so much the better. Here, the randomization test may be of use.

The randomization test addresses the question: Might the two diets really be equally effective and the apparent superiority of the improved diet be attributed to chance variability? The success of the test depends on the relevance of the interpretation it requires for the notions of "equally effective diets" and "chance

* David A. Lane is Assistant Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455.

© Journal of the American Statistical Association
September 1980, Volume 75, Number 371
Theory and Methods Section

variability." According to the Fisherian foundation of the test, two treatments can be considered equally effective only if they would each elicit exactly the same response from each experimental unit. The experimenter may, however, be interested in a weaker notion of equality between two treatments: Their distributions for the responses over the experimental units (or over all potential recipients of the treatments) should coincide. For example, a physician may not believe that each cancer patient faces the same prospect for a cure from radiotherapy as from chemotherapy, but he or she still might want to entertain the hypothesis that the overall success rates of the two treatments might be the same. The randomization test would be of no help to the physician.

The way in which "chance variability" enters into his experiment should be carefully explicated by the scientist when he constructs the statistical model he will use for analyzing his results. The randomization test ignores this model and substitutes an alternative relation between chance and the experiment, based on a frequency distribution induced by the physical act that assigns animals to diets. The logical foundation of this relation is challenged by the scientist's second question.

Basu presses this challenge home and denies Fisher's dictum that the physical act of randomization validates the randomization test. I find his argument convincing, and yet it seems to me that the significance probability of $\frac{1}{2}^{15}$ can possess a rhetorical force that tells for the superiority of the improved diet, without reference to the distribution induced by the physical randomization. To explain this, I need to describe certain thoughts that the scientist might have about his experiment.

For each of his 30 animals, the scientist has ideas at the beginning of the experiment about what the animal would weigh at the end, were it fed the standard diet. Although it undoubtedly implies more precision than the scientist could readily supply, think of these ideas as generating 30 standard-diet predictive distributions. One hypothesis about the diets that the scientist might entertain—although we know he does not believe it!—is H_0: Each animal would end up weighing the same under the standard diet as it would under the improved diet. In particular, if H_0 were true, the 30 standard-diet predictive distributions also describe the scientist's ideas about what the animals would weigh if they were fed the improved diet. Now suppose the scientist has paired his 30 animals so successfully that the predictive distributions for the two animals in each pair coincide. Moreover, suppose also that there are no patterns of covariation among his animals such that, if H_0 were true and he knew the outcome of the experiment for some group of pairs, the scientist's conditional predictive distributions for the two animals in each of the remaining pairs would differ. Call this state of knowledge—or lack of it!—about the experimental animals *null neutrality*.

Under H_0 and null neutrality, the scientist's predictive probability that the 15 animals on the improved diet all end up heavier than their partners is $\frac{1}{2}^{15}$. In fact, the joint predictive distributions under H_0 and null neutrality induce the uniform distribution on the set $S = \{(\pm 1, \ldots, \pm 1)\}$, where the ith coordinate of a point in S is $+1$ if d_i is positive and -1 otherwise. So the usual null distribution of the sign test derives from the scientist's predictive distributions under H_0 and null neutrality, without regard to the method of assignment of animal to diet. The null distribution for the Fisher randomization test can also be derived, with somewhat more tedious assumptions about conditional predictive distributions, in terms of the scientist's prior beliefs about the experimental outcome under H_0. Since the scientist's real beliefs about the superiority of the improved diet imply predictive distributions weighted toward large positive values for the d_i's, small values of the significance probability from the randomization test indicate small posterior probability near H_0, if the scientist had assessed null neutrality and fully probabilized the problem—hence the rhetorical if not inferential force of the $\frac{1}{2}^{15}$.

What about the assessment of null neutrality? It is to be regarded as a rough approximation at best; if the scientist is willing to think hard enough, he can of course recognize differences between any pair of animals. Still, if null neutrality holds approximately, so do the conclusions that follow from assuming it, which serve only as guidelines anyway. In this regard, it is not much different from assuming normality in measurement situations. Yet, just as with normality, it is an assessment not to make lightly—and, as I shall argue later, the physical act of randomization can play a role in deciding whether the assessment is appropriate.

To see whether this or the Fisherian interpretation of the statistician's significance probability provides the sounder guidance for the scientist, it is useful to consider some extreme cases. The issue should not be whether these cases occur in practice, but whether the logic that you claim to follow in practice guides you rightly or wrongly when pressed into extremity.

Example 1 (a variant of Basu's biased coin): The scientist achieves a successful pairing—null neutrality seems reasonable. He generates 15 random numbers on the university computer, associates each of these numbers with a distinct pair of animals, and assigns the first animal (first, relative to a list of the animals' cage addresses) in each pair to the improved diet, if the pair's random number is even. It turns out that, in each pair, the animal that received the improved diet ends up heavier.

Just as the scientist is about to write up his results, however, the computer center informs him that because of a faulty program, only about 40 percent of the random digits the generator produced during the experiment were even.

Example 2: Same story, but the scientist knew about the generator's quirk before he chose the numbers.

Example 3: The scientist is not so lucky as in example 1, or perhaps he knows more about his experimental animals: In each pair, he can identify one animal that seems to have more growth potential than the other. This time, the random-number generator is working fine. Surprisingly, all the animals that the scientist judged to have higher growth potential get assigned to the improved diet. And they all end up heavier.

Basu carefully—and, as far as I can see, successfully—argues that Fisher's logic leads to a significance probability different from $\frac{1}{2}^{15}$ for example 2. The same argument must apply to example 1, since Fisher's logic allows the scientist's knowledge of the randomizing mechanism to enter into the analysis only when he writes down a probabilistic model for it—and since this model attempts to represent the mechanism's physical properties, he must use whatever he knows when he analyzes the experiment, not what he thought he knew when he generated the random numbers. The significance probability derived from the Fisherian logic, as discussed by Basu, is singularly unattractive as a measure of evidence, depending as it does on an artifact of the method of listing cages.

Fisher's logic, tied to the random-number generator and imaginary repetitions, cannot fault the calculation of a significance probability of $\frac{1}{2}^{15}$ in example 3. The design of the experiment may be at fault here, and the experiment itself quite uninformative scientifically, but this does not seem to stand in the way—in the Fisherian framework—of analyzing its evidential content by the $\frac{1}{2}^{15}$ significance probability.

Interpreting the statistician's significance probability in terms of the scientist's predictive probability distribution changes this analysis completely. In example 1, the significance probability of $\frac{1}{2}^{15}$ is unchanged by the computer center's information, since the probability refers to the scientist's thoughts at the commencement of the experiment, to which the information is irrelevant. At first sight, the same holds in example 2, since the probability does not refer to the method of assignment of animal to diet. But the scientist is interested in sharing his assessment of null neutrality: He wants his readers to feel the force of his argument, and so his assessment must be theirs. From this point of view, using a biased or arbitrary mode of assignment is to invite suspicion of loading the experiment in the scientist's favor—perhaps unconsciously, as in the famous Lanarkshire milk experiment ("Student" 1931). Randomly assigning animals to diets with public probability $\frac{1}{2}$ is a way of guaranteeing the honesty—to the public and the scientist himself—of his subjective assessment that both animals in a pair had the same standard-diet predictive distribution.

In example 3, the scientist cannot assess null neutrality, and so the significance probability of $\frac{1}{2}^{15}$ does not apply. Here, he can block his pairs according to his predictive distributions for the d_i's, to ensure as informative an experiment as possible. Again, he can employ one or more physical acts of randomization as a check and guarantee of his subjective assessments of these distributions. The experiment can of course be analyzed, but the null distribution for the randomization test will no longer follow Fisher's frequency distribution and will necessarily be somewhat less open to general agreement.

[*Received December 1979.*]

REFERENCE

Student (1931), "The Lanarkshire Milk Experiment," *Biometrika*, 23, 398.

Comment

D.V. LINDLEY*

What is one to do with this paper but applaud it? Another incoherent procedure has its nature clearly displayed. Here is an encore that I have used in class to suggest that the randomization test does not "pass the test of common sense."

The example is artificial in that the experiment is very small, but this has the virtue of simplifying the arithmetic. The same principle holds for a larger and more realistic experiment at the expense of computations that might obscure the essential ideas. Two scientists are to conduct an experiment to compare a treatment, T, thought to improve the yield, with a control, C. Four units are to be used, two each for T and C. The six possible assignments of T and C to the units are

* D.V. Lindley was formerly Head, Department of Statistics and Computer Science, University College London. He is now retired and lives at 2 Periton Lane, Minehead TA24 8AQ, England.

© Journal of the American Statistical Association
September 1980, Volume 75, Number 371
Theory and Methods Section

listed in the first column of the tabulation appearing two paragraphs after this one. The first scientist, A, decides to select one of the six designs at random. The second scientist, B, feels that the first and last designs would be unsatisfactory, because all the treatments and all the controls come together, and therefore selects a design at random from the four remaining. (In practice, as mentioned before, larger sets of designs would be used.) Both A and B carry out their respective randomizations and both come up with the design *TCTC*, in the second row of the tabulation. On implementing the design, both scientists obtain the results 5, 4, 3, 2 shown in the final row of the tabulation. The total for the treated units is 8, that for the control 6, and the effect is measured by the difference, 2. So far the scientists agree, but now see what happens if they use the randomization argument for analysis.

Had the observed values arisen from any other of the designs that might have been used, the differences would have been those listed in the second column of the tabulation. Consider scientist B first. Scientist B excluded the first and last designs, and so the possible differences are (2, 0, 0, −2), of which the first, the one actually obtained, is the largest. Hence the result is significant at 25 percent, because all designs had the same 25 percent chance of being used. Scientist A, however, included the first and last designs in the randomization so must include the differences 4 and −4 that could have arisen by use of them. Of all six differences, 4 is the largest and 2, the one actually observed, the next largest. Hence the chance of the observed difference, or more extreme differences, is 2 out of 6 and the result is significant at 33⅓ percent.

There, then, are two scientists who have performed exactly the same experiment, *TCTC*, obtained exactly the same result, and yet one is quoting a significance level substantially in excess of the other. And the reason for this difference in level is that A contemplated doing experiments that B did not (viz., those in the first and last rows of the tabulation), although, in fact, A did not perform one of these experiments. Expressed slightly differently, the analysis of the results of the experiment depended on what might have been done, but in fact was not done. Certainly in this context, in which the only probability ideas leading to the level are the equal probabilities involved in the random assignment, the argument seems unsatisfactory.

		Designs	Differences
A		TTCC	4
	B	TCTC	2
		TCCT	0
		CTTC	0
		CTCT	−2
		CCTT	−4
Results		5 4 3 2	

The whole concept of A and B reaching substantially different conclusions seems so absurd that the randomization-analysis argument has to be dismissed. There are two defenses: first, that in practice substantial differences (like 25 and 33⅓ percent) are not observed and that the results are typically the same as normal theory. In that case, why not use normal theory? The second defense is that A and B ought to argue differently because B thought that the first and last experiments might be unsatisfactory, whereas A did not. In other words, both scientists had different ideas before the experiment; is it not reasonable that the two scientists should have different ideas afterwards? This argument violates the claim often made for significance tests—that they allow the data to speak for themselves and are not affected by considerations outside the data—and if admitted plays straight into the Bayesian camp, where the ideas of prior information are considered explicitly.

A minor comment is that it is perhaps a little unfair to say that there is "no satisfactory answer to the question: Why randomize?" The work of Rubin (1978) has at least made a substantial contribution to the answer. The answer for me is tied up with what we mean by *random*. (Basu's definition of randomization, in the first section of Section 2, is in terms of randomness.) I suggest that X is random, given H, if X is independent of any A, given H; that is, if $p(A|X, H) = p(A|H)$. The idea is that the generation of X, whether by a random mechanism, or by pseudorandom numbers, is unconnected with anything else. It is thus a subjective notion, in that what you consider random, I might not; though, in practice, we observe a lot of agreement among people. The value of randomization in design may then be illustrated by an experiment to test the efficacy of treatment T in aiding the recovery R of a patient. We require the probability of a patient's recovery were the patient to be given a treatment, $p(R|T, D)$, using data D from a planned experiment. This may differ from $p(R|T, D, A)$, where A is some factor unrecognized by us. (Had it been recognized it could have been planned for in the acquisition of D.) In order to make reasonably sure that our design does not confound the effects of T and A, we may assign treatments at random, that is, independent of A. This does not ensure lack of confounding but reduces its possibility to an acceptable level. Thus prerandomization has a place in coherent analysis: Basu shows that postrandomization is incoherent.

[*Received December 1979.*]

REFERENCE

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

Comment

DONALD B. RUBIN*

Basu's article on Fisher's randomization test for experimental data (FRTED) is certainly entertaining. Although much of the paper is devoted to the thesis that Fisher changed his views on FRTED, apparently the primary point of the paper is to argue that FRTED is "not logically viable." Admittedly, FRTED is not the ultimate statistical weapon, even in randomized experiments, but calling it illogical is rather bizarre.

Basu criticizes FRTED through two primary arguments. His first line of criticism follows from his attack on a nonparametric test labeled in Section 4 as "Fisher's randomization test." But this test was not proposed by Fisher and is not a logical variant of FRTED; consequently, these criticisms are not of FRTED. I believe that Basu agrees with this contention because in concluding this first criticism he states, "Where is the physical act of randomization in the Fisher randomization test? . . . We should recognize the fact that in Section 21 of *Design of Experiments* (1935) Fisher was not really concerned with the particular test situation that we have discussed in the previous section." Basu's second line of criticism of FRTED takes the form of a discussion between a statistician and a scientist; I find this discussion so confused that it is easier for me to challenge the argument indirectly by clearly describing FRTED than directly by correcting particular misconceptions.

In the paired comparison experiment, let Y_{ij} be the response of the ith unit ($i = 1, \ldots, 2n$) if exposed to treatment j ($j = 1, 2$), where $Y = \{Y_{ij}\}$ is the $2n \times 2$ matrix of values of Y_{ij}. The assumption that such a representation is adequate may be called *the stable unit-treatment value assumption:* If unit i is exposed to treatment j, the observed value of Y will be Y_{ij}; that is, there is no interference between units (Cox 1958, p. 19) leading to different outcomes depending on the treatments other units received and there are no versions of treatments leading to "technical errors" (Neyman 1935). If Y were entirely observed, we could simply calculate the effect of the treatments for these $2n$ units; for example, $Y_{i1} - Y_{i2}$ would be an obvious measure of the effect of treatment 1 versus treatment 2 for the ith unit, and the average value of $Y_{i1} - Y_{i2}$ would be a common measure of the typical effect of treatment 1 versus treatment 2 for these $2n$ units. Because each unit can be exposed to only one treatment, we cannot observe both Y_{i1} and Y_{i2}, and so we will have to draw inferences about the unknown values of Y from observed values of Y.

Let $T = (T_1, \ldots, T_{2n})$ be the indicator for treatment received: $T_i = 1$ if the ith unit received treatment 1 and $T_i = 2$ if the ith unit received treatment 2; if $T_i = 1$, Y_{i1} is observed and Y_{i2} is missing, whereas if $T_i = 2$, Y_{i2} is observed and Y_{i1} is missing. In order to avoid confusion about the inferential content of indices, suppose that the unit indices i are simply a random permutation of $(1, \ldots, 2n)$. The pairing of the units in the paired comparison experiment will be represented by X, where $X_i = 1$ for the two units in the first pair, ..., and $X_i = n$ for the two units in the nth pair. Other characteristics of units can be coded in other variables, but for simplicity assume for now that only values of Y, X, and T will be used for drawing inferences, where Y is partially observed and both X and T are fully observed.

Both randomization and Bayesian inferences for unobserved Y values require a specification for the conditional distribution of T given (Y, X), say $\Pr(T|Y, X)$. The physical act of randomization in the experiment (e.g., the physical act of haphazardly pointing to a starting place in a table of random numbers) is designed to ensure that all scientists will accept the specification $\Pr(T|Y, X) = \Pr(T|X)$. In the paired comparison experiment,

$$\Pr(T|X) = \begin{cases} 0 & \text{if } T_i = T_j \text{ for any } i \neq j \text{ s.t. } X_i = X_j \\ 2^{-n} & \text{otherwise.} \end{cases} \quad (1)$$

If treatments are assigned using characteristics Z of the units that are correlated with Y (the scientist's confessed experiment at the end of Sec. 5), then $\Pr(T|Y, X) = \Pr(T|X)$ would generally not be acceptable. For example, if treatment assignments are determined by tossing biased coins where the bias favors the first unit in each pair receiving treatment 1 (Z = order of unit in pair), then whether $\Pr(T|Y, X) = \Pr(T|X)$ is generally acceptable depends on the scientific view of the partial correlation between Z and Y given X; if the order "does not seem to have much relevance," then $\Pr(T|X, Y) = \Pr(T|X)$ may be plausible with (1) as the accepted specification for $\Pr(T|Y, X)$. Of course, even if unit order is randomly assigned within pairs,

* Donald B. Rubin is Senior Statistical Research Adviser, Educational Testing Service, Princeton, NJ 08541.

© Journal of the American Statistical Association
September 1980, Volume 75, Number 371
Theory and Methods Section

one could decide to record its values and use $\Pr(T|X, Z)$ to draw inferences; this is analogous to recording the random numbers used to assign treatments and observing that given them no randomization took place (i.e., $\Pr(T|X, Z) = 1$ for one value of T and 0 for all other values of T). In order to make sensible use of FRTED, we cannot condition on numbers accepted a priori to be unrelated to Y.

Suppose that we wish to consider the hypothesis H_0 that $Y_{i1} = Y_{i2}$ for all i, or any other sharp null hypothesis such that given H_0 and the observed values in Y, all values of Y are known. Under H_0 and accepting specification (1), the difference in observed averages $\bar{y}_d = \sum Y_{i1}(2 - T_i)/n - \sum Y_{i2}(T_i - 1)/n$, or any other statistic, has a conditional distribution given Y and X consisting of 2^n equally likely known values. Because the expectation of $\bar{y}_d$ over this distribution is zero, values of $\bar{y}_d$ far from zero are a priori considered to be more extreme than values near zero. The proportion of possible values as extreme or more extreme than the observed value of $\bar{y}_d$, that is, the significance level of FRTED is not a property solely of the data and the null hypothesis but also of the statistic and the definition of extremeness of the statistic. If the observed value of $\bar{y}_d$ is extreme (e.g., if the significance level is less than 1 in 20), then we must believe that

1. H_0 is false with the result that the treatments have an effect; or
2. $\Pr(T|Y, X) = \Pr(T|X)$ is false with the result that the 2^n values of $\bar{y}_d$ are not a priori equally likely; or
3. An a priori unusual (extreme) event took place.

The physical act of randomization is designed to rule out option 2 and consequently leave us believing either that an a priori unusual event has taken place or that H_0 is false.

I see nothing illogical about the FRTED; it is relevant for those rare situations when a purely confirmatory test of an a priori sharp hypothesis is to be made using an a priori defined statistic having an associated a priori definition of extremeness. On this point, I find myself in total agreement with the following statement of Brillinger, Jones, and Tukey (1978, p. F–1):

> If we are content to ask about the simplest null hypothesis, that our treatment ("seeding") has absolutely no effect in any instance, then the randomization, that must form part of our design, provides the justification for a randomization analysis of our observed result. We need only choose a measure of extremeness of result, and learn enough about the distribution of this result
> - for the observed results held fixed
> - for re-randomizations varying as is permitted by the specification of the designed process of randomization.
>
> If $p\%$ of the values obtained by calculating as if a random re-randomization had been made are more extreme than (or equally extreme as) the value associated with the actual randomization, then $p\%$ is an appropriate measure of the unlikeliness of the actual result.
>
> Under this very tight hypothesis, this calculation is obviously logically sound.

Of course, there are limitations of FRTED of which Fisher was well aware. For example, the null hypothesis that $Y_{i1} = Y_{i2}$ for all i may not be very realistic; when Neyman (1935) criticized the FRTED for Latin Squares, Fisher (1935a) replied:

> [The null hypothesis that "the treatments were wholly without effect"] may be foolish, but that is what the Z-test [FRTED] was designed for, and the only purpose for which it has been used . . . Dr. Neyman thinks that another test would be more important [one for the average treatment effect being zero]. I am not going to argue that point. It may be that the question which Dr. Neyman thinks should be answered is more important than the one I have proposed and attempted to answer . . . I hope he will invent a test of significance, and a method of experimentation, which will be as accurate for questions he considers to be important as the Latin Square is for the purpose for which it was designed.

More complicated questions, such as those arising from the need to adjust for covariates brought to attention after the conduct of the experiment, simultaneously estimate many effects, or generalize results to other units, require statistical tools more flexible than FRTED. Such tools are essentially based on a specification for $\Pr(Y|X, Z)$, where now Y refers to outcome variables in general, X refers to blocking and design variables, and Z refers to covariates. Fisher (1935a) was certainly willing to specify particular distributional forms for data in experiments, and I believe that he was simply advocating such an attack whenever justified in his "astonishing short section on nonparametric tests in the seventh edition of *DE*." This desire to condition on all relevant information is obviously very Bayesian.

I believe (Rubin 1978) that Bayesian thinking, which requires specifications for both $\Pr(T|Y, X, Z)$ and $\Pr(Y|X, Z)$ and draws inferences conditional on all observed values, provides, in principle, the most effective framework for inference about causal effects. Other statisticians view the specification $\Pr(Y|X, Z)$ as something to be avoided in principle: "For crucial comparisons . . . the appropriate role for the classical kind of parametric analysis would seem to be confined to assistance in the selection of the test statistics to be used . . . in a randomization analysis" (Brillinger, Jones, and Tukey 1978, p. F–5). Using the test statistic (in conjunction with the null hypothesis and definition of extremeness) to summarize all scientific knowledge relevant for data analysis seems to be unduly restrictive. Although much care is needed in applying Bayesian principles because of the sensitivity of inference to the specification $\Pr(Y|X, Z)$, the increased flexibility and directness of the resulting inferences make the Bayesian approach scientifically more satisfying.

On this point, perhaps Basu and I are actually in substantial agreement. FRTED cannot adequately handle the full variety of real data problems that practicing statisticians face when drawing causal infer-

ences, and for this reason it might be illogical to try to rely solely on it in practice.

[*Received December 1979.*]

REFERENCES

Brillinger, D.R., Jones, L.V., and Tukey, J.W. (1978), "The Role of Statistics in Weather Resources Management," Report of the Statistical Task Force to the Weather Modification Advisory Board.

Cox, D.R. (1958), *Planning of Experiments*, New York: John Wiley & Sons.

Fisher, R.A. (1935a) (7th ed. 1960), *The Design of Experiments*, Edinburgh: Oliver and Boyd.

——— (1935b), Discussion of "Statistical Problems in Agricultural Experimentation" by J. Neyman, *Journal of the Royal Statistical Society*, II, 2, 154–180.

Neyman, J. (1935), "Statistical Problems in Agricultural Experimentation," *Journal of the Royal Statistical Society*, II, 2, 107–154.

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

Rejoinder

D. BASU

Let me begin by thanking Hinkley, Lane, Lindley, Rubin, and my good friend Kemp for their many interesting comments. I also offer my apologies to them for my inability, because of an eye condition needing surgical treatment, to read the discussions for myself. They were read out to me, and so I may have missed out on some of the many issues raised. I thank Carlos Pereira for his help in putting together this reply.

Rubin wonders about the relevance of the material discussed in Section 4. Let me explain why I challenged the Fisher nonparametric test—the first nonparametric test by many years, as Fisher (*DE* 1960) put it. The logic of the test is essentially the same as that of the paired-comparison test discussed in Section 6. Both are conditional tests of a very extreme kind. In the nonparametric test, the statistic $(|x_1|, |x_2|, \ldots, |x_n|)$ is held fixed; the δ_i's define the reference set. In the randomization test of Section 6, everything but the design outcome is held fixed. Kempthorne and Folks (1971) labeled the nonparametric test as the Fisher randomization test even though, as I explained at the end of Section 5, the δ_i's cannot really be likened to a set of randomization variables. (Kemp disputes this, but then he disputes almost everything I said.) Each of my difficulties with the nonparametric test also persists with the randomization test. For instance, why must we choose $\bar{x}$ (in Sec. 6, $\bar{d}$) as the test criterion and not the median $\tilde{x}$? With $n = 7$ and each $x_i > 0$, the significance level (SL) works out as 1/128 with $\bar{x}$ as the criterion and as 1/16 with $\tilde{x}$ as the criterion. Neither Kemp or Hinkley answers my question. At one place Kemp mumbles about the central limit theorem, but that is hardly relevant for my sample size. Hinkley makes the curious suggestion that the choice of the test criterion is not a statistical problem. How to justify holding $|x_1|, |x_2|, \ldots, |x_n|$ fixed in the nonparametric test? Why not hold $|\bar{x}|$ fixed instead? In the latter case, the SL is either $\frac{1}{2}$ or 1. In Section 6, when the scientist admitted that he had made a one-toss restricted randomization, the statistician declared the experiment to be uniformative because, for every possible outcome of the experiment, the SL is either $\frac{1}{2}$ or 1. Kemp agrees with the statistician. But Kemp, why? Should we not treat such value-loaded terms like significant or informative with greater respect?

When I said that the Fisher randomization test is not logically viable—Rubin calls the characterization "bizarre" and Kemp, in classical debating style, queries my system of logic—I only meant that the logic of the test procedure is not viable. How else can you characterize a test procedure that falls to pieces when confronted with the slightly altered circumstances of a restricted or unequal probability randomization? I am happy to note that Lane and Lindley agree with me on this point.

My working definition of a Bayesian fellow traveler is one who has trouble in understanding a P value as the level of significance attained by the particular data. Rubin, who claims to be a Bayesian, seems to be quite at home with significance testing. George Box is another notable exception to my working definition.

Let us try to make some sense—please Kemp, do not ask me to define *sense*—of the P value of 2^{-15} in Section 6. Suppose each of the 15 subject pairs is indistinguishable to the scientist. Also suppose that the scientist believes that there is no treatment difference. No doubt then the scientist will be surprised if, at the end of the experiment, he finds that each of the 15 treated subjects gains more weight than the corresponding control subjects. The SL of 2^{-15} may be regarded as

© Journal of the American Statistical Association
September 1980, Volume 75, Number 371
Theory and Methods Section

a measure of this element of surprise. It is a probability (measure of doubt) that existed in the mind of the scientist before the experiment and under the assumed circumstances. As Lane observes, this probability does not depend on the nature of prerandomization. But Kemp, the frequentist, refuses to interpret the SL in terms of such nonexistent belief probabilities.

If the scientist cannot truly distinguish between the subjects in each block, then "Nature has done the randomization for us," says Hinkley, and so he cannot understand the point in all the fuss that I am making. But our scientist, like most scientists, can distinguish between the subjects in each block—one subject is heavier, the other one is older and so on. Mother Nature is asking for a helping hand, and so the scientist must randomize! But the scientist can still distinguish between the subjects in each pair. How can we evaluate his surprise index? So we very sternly tell the scientist, "Randomize and close your eyes!" The scientist randomizes, closes his eyes, but still refuses to be greatly surprised in the end. Because, he says, he knew all along that the improved diet is superior to the standard diet. At this point Kemp will perhaps say, "I am surprised that you can write so much on *surprise* without even defining the term."

Many of my esteemed colleagues believe that postrandomization is a useful statistical device. I know my friend Kemp well enough to say that he is not one among them. He agrees with Fisher, Lindley, and me that postrandomization has no place in scientific thinking. But, today, fighting for every inch of the ground, Kemp is trying to prove me wrong even on this issue. Perhaps one can play better poker by wearing a mask, making hand signals instead of using one's vocal chords, and carrying a randomizer hidden in one's pocket. But does Kemp really think that our scientist is engaged in something like a poker game against Mother Nature? Why does he not advise the scientist also to wear a mask?!

I have no objection to prerandomization as such. Indeed, I think that the scientist ought to prerandomize and have the physical act of randomization properly witnessed and notarized. In this crooked world, how else can he avoid the charge of doctoring his own data? In order to make the device a superior cosmetic agent it may be necessary to make the extent of prerandomization sufficiently wide. In Basu (1978) I have mentioned a few noncosmetic uses of the prerandomization device.

Lindley agrees wholeheartedly with my criticisms of the Fisher randomization test. But, disagreeing with me on what he calls a "minor point," he suggests that there may be a place for randomization in a subjective Bayesian theory of statistics. All I know is that L.J. Savage had similar thoughts but he never spelt them out for us. I may have something to say on the Rubin (1978) thesis on another occasion.

Hinkley and Rubin quote from the prefiducial Fisher to dispute me on the randomization test. In the thirties, Fisher knew that the unrestricted, equal probability randomization test closely parallels the traditional test based on the Gaussian law. So Lindley is asking, "Why not use normal theory?" I remember having seen a Fisher quotation (from the prefiducial time) saying that the randomization test provides a logical justification for the parametric tests based on the normal theory. So Kemp is asking us to discard the normal theory and use the randomization logic instead. In Section 21(a) of *DE* (1960), we find Fisher summarily discarding the Kempthorne thesis on experimental designs. Kemp says that no useful purpose can be served by trying to "psychoanalyze" the mind of Fisher. But what purpose does it serve to dismiss much of Fisher's later writings as mere polemics?

I cannot understand what Hinkley is trying to communicate with his comments on the ancillarity of the design outcome. Is it "plain foolish" to regard the design outcome as an experimental constant? Since there are only a finite number of design outcomes, how can one get an "infinitesimal slice" of the sample space by holding the ancillary statistic fixed? As I pointed out, it is the randomization-test argument that rests on an infinitesimal slice of the sample space by holding fixed everything but the design outcome. The Bayesian recommendation is to hold the data fixed and to speculate about the still-variable parameters. When you push the Fisher conditionality argument to the limit, you become a Bayesian.

On the ancillarity issue, Kemp adopts the proverbial Chinese philosophy of seeing no evil. He is in effect saying, "How can there be an ancillary statistic when there is no probabilistic statistical model and, therefore, no parameters?" I have no difficulty in recognizing the 60 parameters $\omega = \{(x_i, y_i): i = 1, 2, \ldots, 30\}$ in the scientist's diet problem—x_i and y_i are, respectively, the would-be treatment and control responses of subject i at the planning stage of the experiment. Let us suppose that the scientist's parameter of interest is $\theta = \bar{x} - \bar{y}$. Consistent with his prior opinion ξ on ω, the scientist has a prior opinion η on θ. After the experiment, the scientist, having observed 15 of the x_i's, and the complementary set of 15 y_i's, must have drastically revised his prior opinion ξ to a new opinion ξ^*. Consistent with ξ^*, the scientist has then an opinion η^* on the parameter of interest θ.

According to DeFinetti, probability, like beauty, exists only in the mind; it is a formal representation of opinion on parameters. The subjective Bayesian thesis on statistics deals with the process of opinion changes in the very limited context of what we may call statistical parameters. The Bayesian thesis appears to me to be coherent and pertinent to the real issues of scientific inference. That the Bayesian paradigm is useful is slowly gaining recognition. Fuller recognition will take time. But by then it will perhaps be time for us to move on to a more useful paradigm.

When it comes to changing one's opinion on a scientific paradigm, the mind of a stubborn scientist—for that matter, the minds of a whole community of trained scientists—certainly does not, perhaps cannot, follow any logic. In his *Scientific Autobiography and Other Papers* (1949, pp. 33–34) Max Planck wrote, "A new scientific truth does not triumph by convincing the opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it." It rarely happens that Saul becomes Paul.

[*Received March 1979. Revised March 1980.*]

REFERENCE

Planck, Max (1949), *Scientific Autobiography and Other Papers*, New York: Greenwood Press.