Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment
Author(s): Donald B. Rubin
Source: *Journal of the American Statistical Association*, Vol. 75, No. 371 (Sep., 1980), pp. 591-593
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association
Stable URL: http://www.jstor.org/stable/2287653
Accessed: 23-08-2017 18:08 UTC

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, $\ldots$, 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 $\Pr(T \mid Y, X)$. 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 $\Pr(T \mid Y, X) = \Pr(T \mid X)$. In the paired comparison experiment,

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

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 $\Pr(T \mid Y, X) = \Pr(T \mid X)$ 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 $\Pr(T \mid Y, X) = \Pr(T \mid X)$ 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 $\Pr(T \mid X, Y) = \Pr(T \mid X)$ may be plausible with (1) as the accepted specification for $\Pr(T \mid Y, X)$. 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 $\Pr(T \mid X, Z)$ 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., $\Pr(T \mid X, Z) = 1$ 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. $\Pr(T \mid Y, X) = \Pr(T \mid X)$ 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 $\Pr(Y \mid X, Z)$, 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 $\Pr(T \mid Y, X, Z)$ and $\Pr(Y \mid X, Z)$ 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 $\Pr(Y \mid X, Z)$ 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 $\Pr(Y \mid X, Z)$, 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 inferences, 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