does not apply to the case we are considering here. From these explanations it follows that it would be safe to adopt the following definitions. By the term "true value" of the difference of the yields of two varieties, sown on κ selected plots, we mean a
[41]

number Δ associated with the difference of the observed partial averages $X_i - X_j$ in such a way that the probability P_t of preserving the inequality

$$|X_i - X_j - \Delta| < t\sigma_{x_i - x_j}$$

is greater than

$$1 - \frac{1}{t^2}$$

for all $t > 0$.

We can determine empirically that the difference of partial averages of the plots sampled shows a fair agreement with the Gaussian law distribution. This encourages us to name the true difference in yields of two varieties a number δ associated with the difference of the corresponding partial averages, under the condition that the probability of preserving the inequality

$$T_1 < X_i - X_j - \delta < T_2$$

equals

$$\frac{1}{\sigma'_{x_i - x_j}\sqrt{2\pi}} \int_{T_1}^{T_2} \exp\left(-\frac{t^2}{2\sigma'^2_{x_i - x_j}}\right) dt,$$

where,

$$\sigma'^2_{x_i - x_j} = \frac{m - \kappa}{m(\kappa - 1)}\left[\sigma_i^2 + \sigma_j^2 + \frac{2\kappa r}{m - \kappa}\sigma_i\sigma_j\right]$$

and $T_1 < T_2$ are arbitrary numbers. [A misprint (or inconsistency) in the preceding equation has been eliminated; cf. formulas (16) and (17).]

We should remember, however, that this definition is not properly justified.

Of course everything that has been said about the comparison of varieties applies to the comparison of fertilizers.
[42]

Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies

Donald B. Rubin

Dorota Dabrowska and Terry Speed are to be most warmly thanked for bringing this fundamentally important but previously recondite early work of Jerzy Neyman to the attention of the statistical community. It is an honor to be asked to discuss this document, which reinforces Neyman's place as one of our greatest statistical thinkers and clarifies the debt all modern statisticians interested in causal inference owe to Jerzy Neyman. There are several specific contributions in this article (hereafter referred to as Neyman, 1923) that I feel are particularly noteworthy. To delineate these for my discussion, I first provide a brief summary using a mix of Neyman's notation and more standard current notation. I then discuss Neyman's original definition of causal effects in randomized experiments, extensions of it to experiments with interference between units and versions of treatments, and further extensions to observational studies. Three other contributions in Neyman (1923) are also analyzed: his proposal for the completely randomized experiment, his proposal for using repeated-sampling evaluations over randomization distributions, and his specific results on variance estimation in the completely randomized experiment. Throughout, I attempt to relate these contributions of Neyman's to proceeding and contemporary work of R. A. Fisher and others, and to subsequent work, including my own cited in the Dabrowska and Speed introduction. My conclusions regarding the relationship of Neyman (1923) to other work are briefly summarized in the final section.

Donald B. Rubin is Professor and Chairman, Harvard University, Department of Statistics, Science Center, 1 Oxford Street, Cambridge, Massachusetts 02138.

1. AN OVERVIEW OF NEYMAN (1923)

Neyman begins with a description of a field experiment with m plots on which v varieties might be applied: "$\cdots$ U_{ik} is the yield of the ith variety on the kth plot"; U_{ik} is a "potential yield" (Neyman's term) not an observed yield because i indexes all varieties and k indexes all plots, and each plot is exposed to only one variety. Throughout, the collection of poten-

tial outcomes, $\mathbf{U} = \{U_{ik}: i = 1, \cdots v; k = 1, \cdots, m\}$ is treated as a priori fixed but unknown. The "best estimate" of the yield from the ith variety in the field is the average potential outcome over all m plots,

$$a_i = \sum_{k=1}^{m} U_{ik}/m.$$

Neyman calls a_i the best estimate because of his concern with the definition of "true yield," something that he struggled with again in Neyman (1935), as I discuss in Section 3.

Neyman then describes an urn model for determining which variety each plot receives; this model is stochastically identical to the completely randomized experiment with $n = m/v$ plots exposed to each variety. (I use n here rather than Neyman's κ to avoid confusion with Neyman's use of k to index plots.) He notes the lack of independence implied by this restricted sampling of treatments without replacement (i.e., it is impossible for one plot to receive more than one variety, and exactly n plots are exposed to each variety), and he goes on to note that certain formulas for this situation that have been justified on the basis of independence and/or the Gaussian Law (i.e., treating the U_{ij} as normal random variables given some parameters) need more careful consideration. Neyman's nonliteral oral translation to Reid (1982, page 45) is:

> "'Using the method of mathematical expectations I make an effort to solve the problem of the dependence of the expected precision of the experiment on the number of plots in the fields and the number of replications. As far as I know, this problem has not been properly treated thus far.'"

The derivation of the "repeated-sampling randomization-based" expectations and variances of sample estimates then follows. Specifically, with $\mathbf{U}$ fixed, let $E[\cdot \mid \mathbf{U}]$ and $V[\cdot \mid \mathbf{U}]$ indicate the expectation and variance over all possible assignments of v varieties to the $m = nv$ plots with n plots receiving each variety. This description and notation are mine, not Neyman's, and so deserve explanation. First, I use the phrase "repeated-sampling randomization-based" to describe this mode of causal inference (Rubin, 1990, 1991b) in order (a) to emphasize that all randomness comes from the randomization distribution (that is, from the urn model with $\mathbf{U}$ fixed)—hence "randomized-based", and (b) to distinguish Neyman's "repeated-sampling" evaluations under a nonnull distribution from Fisher's randomization-based testing of sharp null hypotheses—more of this in Section 6. Second, I explicitly introduce the notation $E[\cdot \mid \mathbf{U}]$ and $V[\cdot \mid \mathbf{U}]$ so that the framework can be extended to handle model-based modes of inference for causal effects, in particular, Bayesian inference as indicated in Section 7.

Now letting x_i be the sample mean of the n plots actually exposed to the ith variety, and analogously for x_j, Neyman shows that

$$E[x_i - x_j \mid \mathbf{U}] = a_i - a_j. \tag{1}$$

Thus, the standard estimate of the effect of variety i versus variety j, $x_i - x_j$, is unbiased (over repeated randomizations on the m plots) for the implied causal estimand, $a_i - a_j$, the average effect of variety i versus variety j across all m plots.

Furthermore, Neyman then shows, expressed in our notation, that

$$V[x_i - x_j \mid \mathbf{U}] = E\left[\frac{s_i^2}{n} + \frac{s_j^2}{n} \;\middle|\; \mathbf{U}\right] - \frac{1}{m} S^2_{(i-j)}, \tag{2}$$

where s_i^2 is the sample variance of the n observed yields under variety i (with divisor $n - 1$), and analogously for s_j^2, and $S^2_{(i-j)}$ is the variance of the m differences $U_{ik} - U_{jk}$ (with divisor $m - 1$).

Thus, the usual estimate of the variance of estimation, $s_i^2/n + s_j^2/n$, is positively biased (over repeated randomizations on the m plots) unless $S^2_{(i-j)} = 0$, that is, unless $U_{ik} - U_{jk}$ is constant for all k (i.e., unless the variety i versus j effect is additive, to use standard current jargon). Generally, $S^2_{(i-j)}$ depends on the unknown correlation r between the U_{ik} and the U_{jk}, about which there are no data. Neyman's recommendation is to assume $r = 1$, but he considers the problem worthy of future study since this method of establishing variances "has to be considered inaccurate."

Two asides are relevant here for connecting these conclusions to current practice. First, if the m plots in the experiment are thought of as having been randomly sampled from a target population of $N \gg m$ plots for which average causal effects are to be estimated, then the usual estimate of variance of estimation is unbiased (over repeated random sampling of m plots from N and repeated randomizations of treatments to the m chosen plots). Second, extensions based on finding the variance of the variance estimate and applying a Satterthwaite (1946) approximation can provide a degrees of freedom for the variance estimate as a function of r (see, for example, Snedecor and Cochran (1980, page 97), in simpler contexts).

2. ON NEYMAN'S USE OF POTENTIAL YIELDS TO DEFINE CAUSAL EFFECTS IN EXPERIMENTS

As Dabrowska and Speed suggest, one of the most interesting aspects of Neyman's presentation is his explicit use of the notation U_{ik} to indicate the yield of plot k if exposed to variety i drawn according to the urn scheme. This notation became standard for describing possible outcomes of randomized experiments

(e.g., Pitman, 1937, Welch, 1937, McCarthy, 1939, Anscombe, 1948, Kempthorne, 1952, 1955, Brillinger, Jones and Tukey, 1978, and dozens of other places, often assuming additivity as in Cox, 1956, and sometimes being used quite informally as in Freedman, Pisani and Purves, 1978, pages 456–458[1]). An elaboration with "technical errors" appears in Neyman's (1935) discussion of the randomized block and Latin square experiments, although its primary use there is to define a null hypothesis of zero average effects as an alternative to Fisher's null hypothesis of absolutely no effects. Joan Fisher Box (1978, page 263) supports the view that Neyman's model was new, perhaps even to Fisher in 1935, calling it a "novel mathematical model for field experiments." According to Reid (1982, page 45), Neyman himself agrees that the 1923 model was new:

> Neyman has always deprecated the statistical works which he produced in Bydogszcz [which is where Neyman (1923) was done], saying that if there is any merit in them, it is not in the few formulas giving various mathematical expectations but in the construction of a probabilistic model of agricultural trials which, at that time, was a novelty.

Nevertheless, looking back before the twentieth century, we can certainly find seeds of this definition of causal effects among both experimenters and philosophers. For example, Cochran (1978) discusses the great English agronomist, Arthur Young:

> A single comparison or trial was conducted on large plots—an acre or a half acre in a field split into halves—one drilled, one broadcast. Of the two halves, Young (1771) writes: "The soil is exactly the same; the time of culture, and in a word every circumstance equal in both."

It seems clear in this description that Young viewed the ideal pair of plots as being identical, so that the outcome on one plot of drilling would be the same as the outcome on the other of drilling, and likewise for broadcasting, implying that the difference between drilling and broadcasting on either is the causal estimand for each.

Nearly a century later, Claude Bernard, a renowned experimental scientist and medical researcher wrote (Wallace, 1974, page 144):

> The experiment is always the termination of a process of reasoning, whose premises are observation. Example: if the face has movement, what is the nerve? I suppose it is the facial; I cut it. I cut others, leaving the facial intact—the control experiment.

Also in the late nineteenth century, the philosopher John Stuart Mill, when discussing Hume's views offers (Mill, 1973, page 327):[2]

> If a person eats of a particular dish, and dies in consequence, that is would not have died if he had not eaten of it, people would be apt to say that eating of that dish was the source of his death.

Furthermore, Fisher (1918, page 214) wrote the following:[3]

> If we say, "This boy has grown tall because he has been well fed," we are not merely tracing out the cause and effect in an individual instance; we are suggesting that he might quite probably have been worse fed, and that in this case he would have been shorter.

And later, in a letter to Gosset ("Student") reported in "Student" (1923, page 283), Fisher wrote:

> Recognising that not only differences of variety but differences in the conditions of the trials may have affected the yields, we may obtain an estimate of what the variability would be if the conditions of any one trial could be replicated in a number of experiments with the same variety, provided the following simple assumptions hold good. The yield obtained in any experiment is the sum of three quantities, one depending only on the variety; a second, depending only on the 'trial'; and a third, which may be regarded as the 'experimental error' varying independently of variety and trial in a normal distribution about zero with a standard deviation which it is desired to estimate.

Although Fisher's subsequent notation did not explicitly indicate a potential outcome for each plot-variety combination, there are certainly similarities to Neyman's 1923 and 1935 formulations.

Consequently, Neyman's notation seems to have formalized ideas that were relatively firmly in the minds of some experimenters, philosophers and scientists prior to 1923. It is, without a doubt, an extremely important contribution since it allows causal effects (such as $U_{ik} - U_{jk}$) and causal estimands (such as $a_i - a_j$) to be defined without reference to any particular probability model for the data.

---

[1] *My thanks to Neal Thomas for alerting me to this last selection.*

[2] *My thanks to Leland Neuberg for bringing this quotation to my attention.*

[3] *My thanks to Paul Holland, with an assist to Arthur Dempster, for pointing out this quotation, which also appears in Dempster (1990).*

3. EXTENSIONS TO EXPERIMENTS WITH INTERFERENCE BETWEEN UNITS AND VERSIONS OF TREATMENTS

The conceptualization using U_{ik} to represent all potential yields in a completely randomized experiment might not be adequate. Neyman was aware of this in 1923 since he referred to "true yields," which he formalized through the use of "technical errors" in 1935. The true yield of plot k with variety i is the mean across infinite hypothetical replications of the current experiment "...without any change of vegetative conditions or of arrangement..." (Neyman, 1935, page 110). Technical errors are the differences between the particular observable yields in this experiment and the "true yields" and thus have mean zero over the hypothetical replications. They are "...due solely to the inaccuracy of the experimental technique, the vegetative conditions in all our hypothetical experiments being the same" (Neyman, 1935, page 110). Furthermore, "It may be easily assumed that... [the technical error on one plot] is independent of [the technical error] corresponding to some other plot, and is varying normally about zero" (Neyman, 1935, page 114). I have never been able to decide what Neyman really meant by his hypothetical replications. Major problems are that all the replications but one are a priori counterfactual,[4] and Neyman's attendant discussion seems to be devoid of real guidance or implications for practice such as is present in Cox's (1958a, Chapter 1) lucid discussion of these issues.

For example, for many years before 1923, agricultural experimenters had used guard rows between neighboring plots treated differently to avoid "interference between units." To characterize precisely the potential yields in the presence of interference, a more revealing notation, which is not a priori counterfactual, is most helpful. Let $\mathbf{W} = (W_1, \cdots, W_m)$ indicate which varieties the m plots receive, and let $Y_k(\mathbf{W})$ be the yield of the kth plot when the m plots are exposed as indicated by $\mathbf{W}$. Before the assignment of varieties, each $Y_k(\mathbf{W})$ for each possible $\mathbf{W}$ is potentially observable, and thus no $Y_k(\mathbf{W})$ is a priori counterfactual. In cases without interference between units, the simpler U_{ik} notation is adequate. Replacing "varieties" with the more general term "treatments" and "plots" with the more general term "units," I call (Rubin, 1980) the assumption that the U_{ik} notation is adequate, the "stable-unit-treatment-value assumption", SUTVA, or simply "the stability assumption." In the case of possible interference between units, the stability assumption is that, for each k and all possible pairs of treatment assignments $\mathbf{W}$ and $\mathbf{W}'$,

$$Y_k(\mathbf{W}) = Y_k(\mathbf{W}') \quad \text{if } W_k = W_k'.$$

That is, under the stability assumption, the yield of the kth plot when exposed to variety $i = W_k$ is the same no matter what varieties the other plots received. With interference between neighboring plots, $Y_k(\mathbf{W})$ and $Y_k(\mathbf{W}')$ can differ even when $W_k = W_k'$ if neighbors of plot k receive different varieties under $\mathbf{W}$ and $\mathbf{W}'$. Interference between units can be a major issue when studying medical treatments for infectious diseases (e.g., malaria, AIDS) or educational treatments given to children who interact with each other.

Variability in outcome due to variability in the efficacy of nominally identical treatments (e.g., coronary bypass surgery) can be handled in an analogous manner. Variation in efficacy of randomly chosen versions of the same treatment is what I think Neyman was trying to capture with his independent technical errors (Rubin, 1986). To incorporate versions of treatments, simply include an additional variable $\mathbf{V} = (V_1, \cdots, V_m)$ so that $(\mathbf{W}, \mathbf{V})$ indicates both the treatments and the versions of the treatments received by all m plots. (In the context of the completely randomized field experiment of varieties, each V_k must be able to take on at least n values since at least n applications of each variety must be available to conduct the experiment.) Then the potential outcomes allowing for both interference and variability in efficacy are $Y_k(\mathbf{W}, \mathbf{V})$, $k = 1, \cdots, m$, which are, again, a priori not counterfactual. The stability assumption is now that, for each k and each possible pair of assignments $(\mathbf{W}, \mathbf{V})$ and $(\mathbf{W}', \mathbf{V}')$,

$$Y_k(\mathbf{W}, \mathbf{V}) = Y_k(\mathbf{W}', \mathbf{V}') \quad \text{if } W_k = W_k'.$$

Experiments with possible carryover effects and other deviations from stability can be similarly handled.

I believe that notation such as this is more satisfying practically than Neyman's 1923 "true yields" or his 1935 "technical errors" because of its direct correspondence to issues of design. Nevertheless, Neyman's notation appears to provide the first explicit definition of nonnull causal estimands in an experiment that is free of specific models, and this is a major contribution.

4. EXTENSIONS TO THE GENERAL CASE INCLUDING OBSERVATIONAL STUDIES

A further limitation of Neyman's original formulation is that it was entirely tied to randomization-based evaluations, and so for a half-century, it was not perceived as being relevant for defining causal effects in observational (i.e., nonexperimental) studies. To be

[4] *A value is counterfactual if it cannot be observed, that is, if it is entirely hypothetical; see Holland (1986) and its discussion for more on this concept in causal inference.*

fruitful, however, this extension to observational studies requires an explicit model for the assignment mechanism, that is, the specification of a model providing the probability of treatment assignment $\mathbf{W}$ given $\mathbf{U}$ (and other variables if relevant), $\Pr(\mathbf{W} \mid \mathbf{U})$. This model arises directly from explicitly viewing the process for observing components of the potential outcomes, $\mathbf{U}$, as a missing data process (Rubin, 1975, 1976, page 581), and allows randomized experiments to be viewed as having "ignorable" missing data mechanisms because $\Pr(\mathbf{W} \mid \mathbf{U})$ is free of $\mathbf{U}$, and observational studies to be modeled as possibly nonignorable because $\Pr(\mathbf{W} \mid \mathbf{U})$ might depend on unobserved components of $\mathbf{U}$.

In this formulation, causal estimands are defined without reference to either a specific model for the data, $\Pr(\mathbf{U})$, or a specific model for the assignment mechanism, $\Pr(\mathbf{W} \mid \mathbf{U})$. As in Rubin (1976, 1978), the assignment mechanism can reflect both the survey sampling of units into the study from a finite population and the assignment of treatments to the units in the study. As a result, the roles of random sampling of units and randomized assignment of treatments are made explicit for both frequentists (randomization-based and model-based) and Bayesians (including direct-likelihood advocates).

In my opinion, the primary conceptual contribution of my work cited by Dabrowska and Speed in their introduction is the coupling of (a) the extension of the experimental potential-outcome notation to observational studies with (b) an explicit model for the assignment mechanism exhibiting possible dependence on all potential outcomes, and the resultant imbedding of both frequency and Bayesian inference in one coherent structure for inferring causal effects in studies of all kinds. I feel that this formulation is relatively subtle. For example, contrast the explicit role played by randomization in this formulation for likelihood-based inference (i.e., for obtaining ignorable assignment mechanisms) with Kempthorne's (1976, page 497) comment on Fisher's contributions to statistics:

> The work of Fisher abounds in curiosities. One which has struck me forcibly is the absence of any discussion of the relationship of Fisher's ideas on experimentation (DOE) to his general ideas on inference (SI). The latter book contains no discussion of ideas of randomization (except for the irrelevant topic of test randomization) which made DOE so interesting and compelling to investigators in noisy experimental sciences. Can the ideas on randomization and on parametric likelihood theory be fused into a coherent whole? I think not.

Prior to my work and others' in the 1970s and 1980s using the "potential outcomes with assignment mechanism" perspective, the standard approach to causal inference in observational studies used one variable to represent the observed outcome and an indicator to represent treatment assignment. Despite the limitations imposed by using this notation (e.g., treatment assignment is correlated with observed outcome even in a completely randomized experiment, except under the null hypothesis), there were many fine contributions on causal inference in observational studies in the half-century following Neyman (1923), such as Peters (1941), Cochran (1965, etcetera, overviewed in Rubin, 1984), Hill (1965), Campbell and Stanley (1966), and Goldberger (1972), to pick only a few from various fields.

Nevertheless, it appears that in some instances the new perspective has fostered a noticeable increase in clarity of thought, exposition and methodology on observational studies. The contrasts between literature on causal inference before and after the influence of the "potential outcomes with assignment mechanism" perspective can be striking, even within the same author. For example, compare Pratt and Schlaifer (1984) with Pratt and Schlaifer (1988), and Heckman (1979) with Heckman (1989); the earlier publications use the observed outcome notation, whereas the later publications, following direct discussion of their previous work by Rosenbaum and Rubin (1984) and Holland (1989), respectively, explicitly adopt the "potential outcomes with assignment mechanism" perspective. Furthermore, the new perspective does seem to be becoming popular in many fields; for example, in addition to the recent references already directly cited here and in the Dabrowska and Speed introduction, and indirectly in those references, consider the very recent Rosenbaum (1987), Robins (1987, 1989), Greenland and Poole (1988), Smith and Sugden (1988), Sugden (1988), Holland (1988a, b, 1989), Dempster (1990), Kadane and Seidenfeld (1990), Sobel (1990), Rubin (1990, 1991a, b), Gelman and King (1991), Efron and Feldman (1991), and their references.

To believe that those who used and accepted Neyman's experimental model understood its extension to observational studies and its role in defining assignment mechanisms would be, I believe, as fallacious as believing that the thinkers quoted in Section 2 had Neyman's formalization in mind prior to 1923. For example, contrast Cox (1958a) writing in the experimental context using the potential outcomes notation, with Cox and McCullagh (1982, Section 6) writing in the observational study context without the benefit of either the potential-outcomes notation or the explicit consideration of an assignment mechanism; for discussion of their problem, "Lord's paradox," using both tools, see Holland and Rubin (1983).

Of course these comments do not mean that the formulation I am advocating for observational studies

is the only one, but rather that it may have had some novel aspects and useful implications, and seems to be becoming more accepted. Nevertheless, there are many respected workers on causal inference from observational data who do not accept this formulation, some claiming that nonexperimental causal inference requires an entirely different conceptualization (e.g., Granger, 1986); also see a variety of selections from Aigner and Zellner (1988).

5. NEYMAN'S PROPOSAL FOR THE COMPLETELY RANDOMIZED EXPERIMENT

I am in full agreement with Scheffé's (1956) description of Neyman's mathematical model as corresponding to the completely randomized experiment, and I also agree with Dabrowska and Speed that the explicit suggestion to use the urn model to physically assign varieties to plots is absent. This latter conclusion, however, is highly influenced by Neyman's attribution of physically randomized experiments to Fisher and his followers (Neyman, 1935, page 109):

> Owing to the work of R. A. Fisher, "Student" and their followers, it is hardly possible to add anything essential to the present knowledge concerning local experiments.... One of the most important achievements of the English School is their method of planning field experiments known as the method of Randomized Blocks and Latin Squares.

The Neyman (1935, page 112) quotation given in the Dabrowska and Speed introduction also attributes randomization to Fisher. Furthermore, Reid (1982, page 45) quotes Neyman to this effect:

> On one occasion, when someone perceived him as anticipating the English statistician R. A. Fisher in the use of randomization, he objected strenuously:
>
> "... I treated *theoretically* an unrestrictedly randomized agricultural experiment and the randomization was considered as a prerequisite to probabilistic treatment of the results. This is not the same as the recognition that without randomization an experiment has little value irrespective of the subsequent treatment. The latter point is due to Fisher, and I consider it as one of the most valuable of Fisher's achievements."

If these statements from Neyman had been replaced by claims of priority based on his 1923 paper, it would have been difficult for me not to have accepted the position that he had independently envisioned physically randomized experiments even without a translation that used the word "randomization."

Certainly, the idea of random assignment seems to have been "in the air" in 1923. "Student" (1923, pages 281–282) writes: "If now the plots had been randomly placed ..." and "... we are as accurate as if we had devoted twice the area to plots randomly arranged." Also, Fisher and MacKenzie (1923, page 473) write:

> Furthermore, if all the plots were undifferentiated, as if the numbers had been mixed up and written down in random order, the average value of each of the two parts [between and within sums of squares] is proportional to the number of degrees of freedom in the variation of which it is compared.

Nearly forty years earlier, Peirce and Jastrow (reprinted in Stigler, 1980, pages 75–83) used physical randomization to create sequences of binary treatment conditions (heavier versus lighter weight) in a repeated-measure psychological experiment.[5] The primary purpose of the randomization was to create sequences such that "any possible psychological guessing of what changes the operator [experimenter] was likely to select was avoided." I detected no suggestion, however, even implicit, that such randomization could play a useful role in the assignment of treatments to nonhuman units.

Despite the early use of physical randomization by Peirce and Jastrow, the allusions to random assignments by "Student" and the mathematical results using the urn-model formulation in Neyman (1923), all writers since 1925, including Neyman, seem to agree that the first explicit recommendation to make physical randomization an integral part of experimentation was in Fisher (1925) closely followed by Fisher (1926). This situation, with its juxtaposition of implicit suggestion and explicit contrary attribution from the same author, emphasizes to me the dangers of overinterpreting, with ebullient and embellished hindsight, early writings of great men.

6. NEYMAN'S PROPOSAL FOR USING REPEATED-SAMPLING EVALUATIONS OVER RANDOMIZATION DISTRIBUTIONS

To my knowledge, this paper represents the first attempt to evaluate, formally or informally, the repeated-sampling properties of statistics over their nonnull randomization distributions, and so I believe this contribution is uniquely and distinctly Neyman's. That Neyman did this prior to his 1935 article is no surprise; in Rubin (1990) I attributed this general mode of inference to Neyman, but lacking Neyman (1923) I referred to Neyman (1934) on random sampling and to Neyman (1935). This mode of inference not only became the standard in survey methodology

[5] *My thanks to Stephen Stigler for calling my attention to this, apparently first, use of randomization in experiments.*

and experimental design, but it is still the standard approach in survey practice as well as the foundation for much of statistical practice in analysis of variance contexts (e.g., the Cornfield and Tukey, 1956, rules for expected mean squares).

Fisher's mode of randomization-based inference in experiments was distinctly different from Neyman's, and, in contrast to Neyman's, had no neat analog for survey practice. Fisher's proposition was to: posit a sharp null hypothesis under which all values are known (e.g., in Neyman's notation $U_{ik} = U_{jk}$ for all i, j pairs and each k); calculate the null randomization distribution of some statistic (i.e., calculate the value of the statistic under the null hypothesis for each possible randomization); locate the observed value of the statistic in its randomization distribution; and, finally, calculate the unusualness (i.e., p-value) of the observed value of the statistic according to some a priori definition of unusualness (i.e., the proportion of the possible values of the statistic as unusual or more unusual than the observed value).

Neyman's prescription offers a general plan for evaluating proposed procedures, whereas Fisher's prescription directly provides distribution-free p-values for sharp null hypotheses. I find the approaches to be complementary.

7. SPECIFIC RESULTS ON VARIANCE ESTIMATION IN THE COMPLETELY RANDOMIZED EXPERIMENT

In the last half-century, many statisticians have repeated the calculations Neyman provides. The most interesting result here is the inherent uncertainty of variance estimation due to the "inestimable" correlation between the $\{U_{ik}\colon k = 1, \cdots, m\}$ and the $\{U_{jk}\colon k = 1, \cdots, m\}$, inestimable because U_{ik} and U_{jk} can never be jointly observed—"the fundamental problem of causal inference" (Holland, 1986). Incidentally, it is only their partial correlation, given observed blocking factors and covariates, that is entirely inestimable.

Given Neyman's stated motivation for deriving equations (1) and (2) (i.e., issues of dependence and normality), I believe it is important to compare his repeated-sampling randomization-based answers for the completely randomized experiment with the corresponding Bayesian answers (following the framework in Rubin, 1978). Specifically, a completely randomized experiment corresponds to an ignorable treatment assignment mechanism, and therefore the posterior distribution of any causal estimand, such as $a_i - a_j$, follows from the specification of a distribution for the data matrix, $\mathbf{U}$. To simplify the comparison of answers, we suppose only two treatments. Then, appealing to de Finetti's theorem, we can write

$$\Pr(\mathbf{U}) = \int \prod_{k=1}^{m} f(U_{1k}, U_{2k} \mid \theta) p(\theta)\, d\theta$$

for some bivariate density $f(\cdot \mid \theta)$ indexed by parameter θ with prior distribution $p(\theta)$. Suppose $f(\cdot \mid \theta)$ is normal with means $\mu = (\mu_1, \mu_2)$, variances (σ_1^2, σ_2^2) and correlation ρ. Then conditional on (a) θ, (b) the observed values of $\mathbf{U}$, $\mathbf{U}_{\text{obs}}$, and (c) the observed value of the treatment assignment, $\mathbf{W}_{\text{obs}}$, we have that the joint distribution of (a_1, a_2) is normal with means

$$\frac{1}{2}\left[x_1 + \mu_1 + \rho \frac{\sigma_1}{\sigma_2}(x_2 - \mu_2)\right],$$

$$\frac{1}{2}\left[x_2 + \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x_1 - \mu_1)\right],$$

variances $\sigma_1^2(1 - \rho^2)/4n$, $\sigma_2^2(1 - \rho^2)/4n$, and zero correlation. To simplify comparison with the repeated-sampling randomization-based answers, now assume large m and a relatively diffuse prior distribution for $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ given ρ. Then the conditional posterior distribution of $(a_1 - a_2)$ given ρ is normal with mean

$$E[a_1 - a_2 \mid \mathbf{U}_{\text{obs}}, \mathbf{W}_{\text{obs}}, \rho] = x_1 - x_2 \tag{3}$$

and variance

$$V[a_1 - a_2 \mid \mathbf{U}_{\text{obs}}, \mathbf{W}_{\text{obs}}, \rho] = \frac{s_1^2}{n} + \frac{s_2^2}{n} - \frac{1}{m}\sigma_{(1-2)}^2, \tag{4}$$

where $\sigma_{(1-2)}^2$ is the prior variance of the differences $U_{1k} - U_{2k}$, $\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho$.

The practical similarity between the Bayesian results (3) and (4) and the repeated-sampling randomization-based results (1) and (2) is striking. From (1) and (3), the correct estimate of the causal effect of variety 1 versus variety 2 is the difference in observed sample means, and from (2) and (4), the usual squared standard error associated with this estimate is too conservative by the amount

$$\frac{1}{m}\sigma_{(1-2)}^2 \doteq \frac{1}{m} S_{(1-2)}^2.$$

Also, if the m plots were randomly sampled from a field of $N \gg m$ plots and the causal estimand were the mean difference between variety 1 and variety 2 across all N plots, this conservatism would vanish from the Bayesian perspective as well as from the Neyman perspective.

Thus, as often occurs with such problems, the Bayesian answer closely parallels the randomization-based answer. A more complete Bayesian derivation would automatically include adjustments for (a) small-

sample effects of variance estimation (i.e., the marginal posterior distribution of $a_1 - a_2$ given ρ is similar to a Behrens-Fisher distribution, which often can be well approximated by a t using a Satterthwaite approximation), and (b) uncertainty due to the unknown correlation (i.e., by integrating over the prior distribution of ρ). Furthermore, simple arguments support the claim that the normal posterior distribution given by (3) and (4) is a good approximation even without the normal assumption for $f(\cdot \mid \theta)$.

8. CONCLUSIONS

Without a doubt, Neyman (1923) is an important, but previously unposted, milestone in statistics. My belief is that the proposal to evaluate procedures over their repeated-sampling randomization-based distributions is uniquely Neyman's. Had it not been for his attributions to the contrary, I would have thought that the proposal to use the physical act of randomization in experimental design was previewed here as well. Finally, with respect to his definition of causal effects, although the underlying implicit definition was relatively common prior to 1923, Neyman certainly appears to be the first to formalize it. However, neither he nor other writers in the next half-century seem to have applied this notation for potential outcomes to observational studies for causal effects, instead using the generally inferior observed-outcome notation, and providing no formal statement of a treatment assignment mechanism exhibiting possible dependence on the potential outcomes. In contrast, in the last dozen years, since the publication of the papers referenced in the Dabrowska and Speed introduction, this framework, with explicit statements of its associated assumptions and explicit modeling of nonrandomized assignment mechanisms, has been applied in a variety of disciplines, often with an attendant increase in clarity. As with Neyman's 1923 formalization, I have no doubt that these refinements were "in the air," and I'm glad to have been a contributor to their exposition and development.

ACKNOWLEDGMENT

This work was supported by NSF Grant SES-88-05433.

ADDITIONAL REFERENCES

AIGNER, D. J. and ZELLNER, A. (1988). *Causality.* Supplement to *J. Econometrics* **39**. North-Holland, Amsterdam.

ANSCOMBE, F. J. (1948). The validity of comparative experiments. *J. Roy. Statist. Soc. Ser. A* **61** 181–211.

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

BRILLINGER, D. R., JONES, L. V. and TUKEY, J. W. (1978). Report of the statistical task force for the weather modification advisory board. *The Management of Western Resources, Vol. II: The Role of Statistics on Weather Resources Management.* Stock no. 003-018-00091-1, U.S. Govt. Printing Office, Washington, D.C.

CAMPBELL, D. T. (1978). Quasi-experimental design. In *International Encyclopedia of Statistics* (J. Tanur and W. Kruskal, eds.) **1** 299–305. MacMillan, New York.

CAMPBELL, D. T. and STANLEY, J. C. (1966). *Experimental and Quasi-Experimental Designs for Research.* Rand McNally, Chicago.

COCHRAN, W. G. (1965). The planning of observational studies of human populations. *J. Roy. Statist. Soc. Ser. A* **128** 234–265.

COCHRAN, W. G. (1978). Early development of techniques in comparative experimentation. In *On the History of Statistics and Probability* (D. Owen, ed.) 2–25. Dekker, New York.

CORNFIELD, J. and TUKEY, J. W. (1956). Average values of mean squares in factorials. *Ann. Math. Statist.* **27** 907–949.

COX, D. R. (1956). A note on weighted randomization. *Ann. Math. Statist.* **27** 1144–1151.

COX, D. R. (1958a). *The Planning of Experiments.* Wiley, New York.

COX, D. R. (1958b). The interpretation of the effects of nonadditivity in the Latin square. *Biometrika* **45** 69–73.

COX, D. R. and MCCULLAGH, P. (1982). Some aspects of analysis of covariance. *Biometrics* **38** 541–561.

DEMPSTER, A. P. (1991). Causality and statistics. *J. Statist. Plann. Inference* **25** 261–278.

EFRON, B. and FELDMAN, D. (1991). Compliance as an explanatory variable in clinical trials. *J. Amer. Statist. Assoc.* **86** To appear.

FISHER, R. A. (1918). The causes of human variability. *Eugenics Rev.* **10** 213–220.

FISHER, R. A. (1925). *Statistical Methods for Research Workers,* 1st ed. Oliver and Boyd, Edinburgh.

FISHER, R. A. (1926). The arrangement of field experiments. *J. Ministry of Agriculture of Great Britain* **33** 503–513.

FISHER, R. A. and MACKENZIE, W. A. (1923). Studies in crop variation. II. The manurial response of different potato varieties. *J. Agricultural Sci.* **13** 311–320.

FREEDMAN, D., PISANI, R. and PURVES, R. (1978). *Statistics.* Norton, New York.

GELMAN, A. and KING, G. (1991). Estimating incumbency advantage without bias. *Am. J. Political Sci.* To appear.

GOLDBERGER, A. S. (1972). Selection bias in evaluating treatment effects: Some formal illustrations. Discussion paper No. 123-72, Institute for Research on Poverty, Univ. Wisconsin, Madison.

GRANGER, C. W. J. (1986). Comment on "Statistics and causal inference" by P. W. Holland. *J. Amer. Statist. Assoc.* **81** 967–968.

GREENLAND, S. and POOLE, C. (1988). Invariants and noninvariants in the concept of interdependent effects. *Scand. J. Work and Environmental Health* **14** 125–129.

HECKMAN, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **47** 153–161.

HECKMAN, J. J. (1989). Causal inference and nonrandom samples. *J. Educ. Statist.* **14** 159–168.

HILL, A. B. (1965). The environment and disease: Association or causation. *Proc. Roy. Soc. Med.* **58** 295–300.

HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–968.

HOLLAND, P. W. (1988a). Causal inference, path analysis, and recursive structural equation models. In *Sociological Methodology* (C. Clogg, ed.) 449–484. Amer. Sociological Assoc., Washington, D. C.

HOLLAND, P. W. (1988b). Comment on "Employment discrimination and statistical science" by A. P. Dempster. *Statist. Sci.* **3** 186–188.

HOLLAND, P. W. (1989). It's very clear. Comment on "Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training" by J. Heckman and V. Hotz. *J. Amer. Statist. Assoc.* **84** 875–877.

HOLLAND, P. W. and RUBIN, D. B. (1983). On Lord's paradox. In *Principals of Modern Psychological Measurement* (H. Wainer and S. Messick, eds.) 3–25. Erlbaum, Hillsdale, N.J.

KADANE, J. B. and SEIDENFELD, T. (1990). Randomization in a Bayesian perspective. *J. Statist. Plann. Inference* **25** 329–346.

KEMPTHORNE, O. (1952). *The Design and Analysis of Experiments.* Wiley, New York.

KEMPTHORNE, O. (1955). The randomization theory of experimental inference. *J. Amer. Statist. Assoc.* **50** 946–967.

KEMPTHORNE, O. (1976). Comment on "On Rereading R. A. Fisher" by L. J. Savage. *Ann. Statist.* **4** 495–497.

MCCARTHY, M. D. (1939). On the application of the z-test to randomized blocks. *Ann. Math. Statist.* **10** 337.

MILL, J. S. (1973). A system of logic. In *Collected Works of John Stuart Mill* **7**. Univ. Toronto Press.

NEYMAN, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *J. Roy. Statist. Soc. Ser. A* **97** 558–606.

NEYMAN, J., with cooperation of K. Iwaskiewicz and St. Kolodziejczyk (1935). Statistical problems in agricultural experimentation (with discussion). *Suppl. J. Roy. Statist. Soc. Ser. B* **2** 107–180.

PETERS, C. C. (1941). A method of matching groups for experiment with no loss of population. *J. Educ. Res.* **34** 606–612.

PITMAN, E. J. G. (1937). Significance tests which can be applied to samples from any populations. III. The analysis of variance test. *Biometrika* **29** 322–335.

PRATT, J. W. and SCHLAIFER, R. (1984). On the nature and discovery of structure (with discussion). *J. Amer. Statist. Assoc.* **79** 9–33.

PRATT, J. W. and SCHLAIFER, R. (1988). On the interpretation and observation of laws. *J. Econometrics* **39** 23–52.

REID, C. (1982). *Neyman from Life.* Springer, New York.

ROBINS, J. M. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *J. Chronic Diseases* (Suppl. 2) **40** 139S–161S.

ROBINS, J. M. (1989). The control of confounding by intermediate variables. *Statist. Med.* **8** 679–701.

ROSENBAUM, P. R. (1987). The role of a second control group in an observational study (with discussion). *Statist. Sci.* **2** 292–316.

ROSENBAUM, P. R. and RUBIN, D. B. (1984). Estimating the effects caused by treatments. Comment on "On the nature and discovery of structure" by J. W. Pratt and R. Schlaifer. *J. Amer. Statist. Assoc.* **79** 26–28.

RUBIN, D. B. (1975). Bayesian inference for causality: The importance of randomization. In *Social Statistics Section, Proceedings of the Amer. Statist. Assoc.* 233–239.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592.

RUBIN, D. B.(1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **7** 34–58.

RUBIN, D. B. (1980). Comment on "Randomization analysis of experimental data: The Fisher randomization test" by D. Basu. *J. Amer. Statist. Assoc.* **75** 591–593.

RUBIN, D. B. (1984). William G. Cochran's contributions to the design, analysis, and evaluation of observational studies. In *W. G. Cochran's Impact on Statistics* (S. Rao and J. Sedransk, eds.) 37–69. Wiley, New York.

RUBIN, D. B. (1986). Which ifs have causal answers? Comment on "Statistics and causal inference" by P. W. Holland. *J. Amer. Statist. Assoc.* **81** 961–962.

RUBIN, D. B. (1990). Formal modes of statistical inference for causal effect. *J. Statist. Plann. Inference* **25** 279–292.

RUBIN, D. B. (1991a). Dose-response estimands: A comment on Efron and Feldman. *J. Amer. Statist. Assoc.* **86**. To appear.

RUBIN, D. B. (1991b). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics* **47**. To appear.

SATTERTHWAITE, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bull.* **2** 110–114.

SMITH, T. M. F. and SUGDEN, R. A. (1988). Sampling and assignment mechanisms in experiments, surveys and observational studies. *Internat. Statist. Rev.* **56** 165–180.

SNEDECOR, G. W. and COCHRAN, W. G. (1980). *Statistical Methods*, 7th ed. Iowa State Univ. Press, Ames.

SOBEL, M. E. (1990). Effect analysis and causation in linear structural equation models. *Psychometrika* **55** 495–515.

STIGLER, S. M. (1980). *American Contributions to Mathematical Statistics in the Nineteenth Century* **2**. Arno Press, New York.

"STUDENT" (1923). On testing varieties of cereals. *Biometrika* **15** 271–293.

SUGDEN, R. A. (1988). The 2×2 table in observational studies. In *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 785–790. Oxford Univ. Press, New York.

WALLACE, W. A. (1974). *Causality and Scientific Explanation: Classical and Contemporary Science* **2**. Univ. Michigan Press, Ann Arbor.

WELCH, B. L. (1937). On the z test in randomized blocks and Latin squares. *Biometrika* **29** 21–52.