R. A. FISHER ON THE DESIGN OF EXPERIMENTS AND STATISTICAL ESTIMATION

0. INTRODUCTION AND OUTLINE

On this occasion of the centenary year of Fisher's birth, my purpose in this talk is to consider the relation between two of Fisher's major contributions: his theory of experimental design and his theory of statistical estimation. It is no coincidence that each of Fisher's three principal books on statistics ends with a substantial chapter on statistical estimation:

Statistical Methods for Research Workers (SMfRW) 1925 — 14th ed. 1970;
The Design of Experiments (DoE) 1935 — 8th ed. 1966;
Statistical Methods and Scientific Inference (SM&MI) 1956 — 3rd ed. 1973.[1]

The thesis of my presentation is that Fisher linked experimental design and estimation through his technical account of (Fisher-) *Information.* In particular, improvements in an experimental design, e.g., better controls, blocking, or other factorial restrictions, may be quantified by an increase in the Information provided by estimates derived from the experimental data.

In Section 1, I sketch Fisher's theory of estimation with an eye on explicating the role Information plays in resolving choices of rival estimates derived from a sample. As an illustration of this approach, I show how Fisher's χ^2-significance test for ordinary contingency tables may be decomposed to reveal that Information justifies maximum likelihood estimation. The same example indicates the importance Fisher placed on the *principle of ancillarity*, to wit: conditioning on ancillary data. In Section 2, I discuss the application of Fisher Information to questions of experimental design, and illustrate how added controls (in matched pairs) lead to data with greater Information. The presentation concludes with a brief discussion of one difficulty in this reconstruction of Fisher's use of Information: the problem of randomization. In estimation it is to be avoided (as a result of the ancillarity principle), yet more

Sahotra Sarkar (ed.), The Founders of Evolutionary Genetics, 23—36.

than anyone else Fisher is responsible for the theory of randomized experimental design. How is this conflict to be resolved?

1. INFORMATION AND FISHER'S THEORY OF ESTIMATION — WITH AN APPLICATION TO 2 × 2 CONTINGENCY TABLES

In his ground-breaking 1922 and 1925 papers on estimation, Fisher offers formal criteria for determining the adequacy of a statistical estimator. In particular, the arguments he gives press for supremacy of maximum likelihood estimation. In rough form, these and later versions of his theory are organized into a hierarchy of criteria, a sequence of ever finer sieves to distinguish among rival estimates. Those which pass the simpler tests (e.g., for "consistency") are subjected to the heightened scrutiny of more refined tests (e.g., for "efficiency"). But what is the goal of estimation? How are the test criteria justified? The key to answering these questions is that for Fisher an estimate is, first of all, a statistic — a reduction of data. That is, an estimate is appraised as a summary of evidence, not as a "guesstimate" of some unobserved quantity.

The first requirement in estimation is Fisher's criterion of consistency. (Fisher-) Consistency of an estimate identifies the quantity (the parameter) about which the summary is directed. It finds its mature formulation (Fisher, 1956, §6.2) as follows:

DEFINITION. A (*Fisher-*) *Consistent Statistic* is a function of the observed frequencies which takes the exact parametric value when for these frequencies their expectations are substituted.

Suppose, for example, the (i.i.d.) N data are categorial, each occupying one of m cells, with probability p_j for the jth cell ($j = 1, \ldots, m$). Denote the observed cell frequencies by a_j/N, where a_j is the jth cell count. Consider a (linear) statistic A of the form $A = \Sigma_j c_j a_j$, for known constants c_j ($j = 1, \ldots, m$). If we substitute for the cell counts their expected values (Np_j), we obtain an estimate $A/N = \Sigma_j c_j p_j$, which (by definition) is a consistent estimate of this (linear) function of the cell probabilities. For instance, it might be that, as with data relating to the genetic linkage between two characteristics, the parameter of interest, θ, satisfies $\theta^2 = \Sigma_j c_j p_j$. Then $\sqrt{(A/N)}$ is a Fisher-consistent estimate of θ. However, the same data may suggest numerous (Fisher-) consistent estimators for the same parameter of interest.

ILLUSTRATION (SMfRW, §53). Corresponding to estimation of a linkage parameter, $0 < \theta < 1$ ($\theta = 0.5$ for independence between the genes), with four cells having respective probabilities of occurrence on each trial: $\{(2 + \theta^2)/4, (1 - \theta^2)/4, (1 - \theta^2)/4, \theta^2/4\}$, then the following three all are Fisher-consistent estimates of θ.

Estimator$_1$: $\sqrt{[(a_1 + a_4 - a_2 - a_3)/N]}$

Estimator$_2$ (the maximum likelihood estimate): the positive solution to

$$a_1/N(2 + \theta^2) + a_4/N\theta^2 = (a_3 + a_4)/N(1 - \theta^2)$$

Estimator$_3$ (the minimum χ^2 estimate): the positive solution to

$$a_1^2/N^2(2 + \theta^2)^2 + a_4^2/N^2\theta^4 = (a_3^2 + a_4^2)/N^2(1 - \theta^2)^2.$$

What is the reason for insisting on Fisher-consistency of an estimate? The answer lies in Fisher's semantics (that is his "theory") of probability. To assert that the data are an (i.i.d.) sample of N according to the parametrized distribution $p_j(\theta)$ is to require (among other conditions) that the "hypothetical" population from which the sample is taken has cell frequencies given by distribution $p_j(\theta)$ ($j = 1, \ldots, m$). Moreover, the parameter is identified by these population quantities. Thus, a Fisher-consistent estimator for a parameter θ meets the quite minimal condition that, when applied to the population itself, the estimator recovers the quantity θ from such an idealized sample. In that sense, the estimator summarizes the (idealized) evidence of the whole population by the parametric quantity of interest.[2]

The next two, closely related criteria by which estimators are assessed involve Fisher's concept of statistical Information. Formally,

DEFINITION. The (expected) *amount of Information* about θ, $I_\theta[x_1, \ldots, x_N]$, in an i.i.d. sample of N from the distribution $p_j(\theta)$ is

$$-\mathrm{E}\left\{\frac{\partial^2}{\partial\theta^2}(\log p_j)\right\}$$

where the expectation is over all possible samples, taken with respect to the distribution $p_j(\theta)$. Information is additive for independent samples, and the Information in a statistic derived from a sample is always bounded above by the Information in the sample as a whole (Fisher, 1925a).

Not surprising, Information serves as a basis in Fisher's theory for discriminating among consistent estimators. Besides consistency, a good estimator T is to be an *efficient* summary of the data from which it is derived. More precisely, with increasing sample size, the ratio of the amount of Information in T_N to the amount of Information in the sample of N from which it is calculated should approach unity:

DEFINITION. Estimator T is (1st order) *efficient* if $\lim_{N \to \infty} I_\theta[T_N] \div I_\theta[x_1, \ldots, x_N] = 1$.

Among the three consistent estimators presented in the gene-linkage illustration (above), the maximum likelihood and the minimum chi-square estimation are efficient. However, the first estimator has an efficiency increasing with the value of the parameter of interest. For example, its efficiency is only about 60% when the two genes are independent ($\theta = 0.5$).

In order to motivate the final criterion for estimates that I will discuss, consider two statistical extremes with respect to the adequacy of a statistic in summarizing a data set.

DEFINITION. A statistic $T(X)$ calculated from a sample x is *sufficient* for the parameter θ provided that $p(X|\, T, \theta) = p(X|\, T)$, independent of θ.

DEFINITION. A statistic $T(X)$ calculated from a sample X is *ancillary* for the parameter θ provided that $p(T|\, \theta) = p(T)$, independent of θ.

Either by Bayesian or likelihood principles, a sufficient statistic conserves all the relevant evidence (about θ) in the sample from which it is derived, whereas an ancillary statistic is irrelevant to θ. From the standpoint of (Fisher) Information, when $T(X)$ is sufficient for θ, $I_\theta[X] = I_\theta[T]$; T preserves all the relevant evidence in the sample. Likewise, when T is ancillary for θ, $I_\theta[T] = 0$; there is no relevant information contained in an ancillary quantity. Thus, in the case of ancillary quantity T, the statistical analysis may be carried out given T. For example, often it is assumed that sample size, N, is chosen independent of the unknown quantity of interest. Then, N is ancillary and the analysis proceeds with N a known constant.[3]

Much of Fisher's attention (particularly in his 1934 paper, 'Two New Properties of Mathematical Likelihood') is devoted to establishing the

supremacy of maximum likelihood estimation [m.l.e.]. At first (1922) he thought the m.l.e. is sufficient. He weakened his claim (1935, §3) to say that the m.l.e. is sufficient whenever a sufficient statistic exists, and that by conditioning on an ancillary statistic the m.l.e. preserves the greatest quantity of Information that can be summarized in a statistic.[4] In order to see what Fisher had in mind, recall his somewhat controversial treatment of contingency tables.

ILLUSTRATION. For simplicity, let us attend to the elementary case of a 2×2 table. Consider data from flips of two coins, summarized in the table below, where we are concerned about the hypothesis that the coins have a common bias.

2×2 TABLE

	heads	*tails*	
coin-1	a	b	$a + b = n_1$ flips
coin-2	c	d	$c + d = n_2$ flips
	$a + c$ heads	$b + d$ tails	$n_1 + n_2 = N$ flips total.

Let θ_0 be an hypothesized value for the common bias of the two coins. Then the χ^2-test for independence (with 2-degrees of freedom) is just the sum of the two, separate 1-degree of freedom χ^2-tests for the two samples (of sizes n_1 and n_2, respectively) about the coins:

$$\chi^2_2 = (a - \theta_0 n_1)^2/\theta_0 n_1 + (b - [1 - \theta_0]n_1)^2/[1 - \theta_0]n_1 + (c - \theta_0 n_2)^2/\theta_0 n_2 + (d - [1 - \theta_0]n_2)^2/[1 - \theta_0]n_2.$$

Fisher's controversial proposal for the χ^2-independence test when θ_0 is unknown is to substitute the m.l.e. under the "null" hypothesis, to replace θ_0 with the quantity $(a + c)/N$, resulting in a χ^2 test with 1 degree of freedom. What connection is there between this use of the maximum likelihood estimate in the χ^2 test and the importance that estimates conserve Information? The answer is both subtle and rather surprising.

Significance tests offer a rudimentary form of (Fisherian) statistical analysis against more sophisticated Fisherian tools, e.g., using the likelihood function or, in the special circumstances where it applies, using fiducial inference to relate data to hypotheses. Significance tests

are rudimentary in their conclusion — "Either a rare event has occurred or the 'null' hypothesis is false." However, to offset this weakness, Fisher argues that significance tests may be performed even when the space of alternative hypotheses is vaguely specified — unlike the conditions for likelihood or fiducial reasoning.

What makes an outcome "rare"? One scheme for making sense of Fisher's idea (well stated by Cramér, 1946) is to introduce a discrepancy ranking **D** on the sample space of possible outcomes Ω, **D**: $\Omega \rightarrow \Re$. The intended meaning is that outcomes with higher discrepancy are "rarer" under the null hypothesis. Then, the significance level attained on a trial is the probability (given h_0) of obtaining a discrepancy at least as great as that observed.

With multinomial data, Fisher (SMfRW, §21.02) favored the "exact test," where discrepancy is inversely related to the probability of cell counts. Call this the probabilistic discrepancy ranking, D_p. That is, with m cell counts, $a_j (j = 1, \ldots, m)$, $\Sigma_j a_j = N$, then

$$D_p\{a_j\} = [(N! \div \Pi_j a_j!)\, \Pi_j p_j^{a_j}]^{-1}.$$

Asymptotically, for increasing sample size, the probabilistic discrepancy ranking agrees with the χ^2-discrepancy ranking (on m-1 degrees of freedom), D_{χ^2}; where outcomes are given a D_{χ^2} discrepancy according to their χ^2 values. Hence, with categorical data, χ^2 gives a convenient approximation to Fisher's "exact" significant test.

The controversial aspects of Fisher's treatment of independence tests in contingency tables stems from his claim that the analysis should take the marginal totals as given. That is, Fisher treats the 2×2 tables as though the three quantities, N, n_1, and $a + c$ are ancillary.[5] In the illustration with the two biased coins, it is commonplace to assume that the sample sizes (n_1 and n_2, hence also N) are irrelevant to inference from the data. The assertion that $a + c$, too, is uninformative about the null hypothesis (of independence) is without foundation. Clearly, it is false for extreme cases, e.g., when $a + c = 0$.

However, if we grant Fisher's assumption, that the marginal totals give no relevant information in the test of the null hypothesis (that the coins are equally biased), then the "exact" test assumes a convenient form (independent of the value θ_0 of the common bias). Then, given the four marginal totals, $D_p(\{a, b, c, d\}) = [a!b!c!d!]$. How does this compare to the D_{χ^2} discrepancy ranking for the same null hypothesis?

In particular, what of Fisher's use of the m.l.e. in computing the 1-degree of freedom χ^2?

The answer is contained in a decomposition of χ^2. Recall,

$$\chi_2^2 = (a - \theta_0 n_1)^2/\theta_0 n_1 \quad + \quad (b - [1 - \theta_0]n_1)^2/[1 - \theta_0]n_1$$
$$+ (c - \theta_0 n_2)^2/\theta_0 n_2 \quad + \quad (d - [1 - \theta_0]n_2)^2/[1 - \theta_0]n_2.$$

By some simple algebra,

$$= [\theta_0 - (a + c)/N]^2 \, N \div \theta_0(1 - \theta_0)$$
$$+ (n_1 n_2 \div N)\{a/n_1 - c/n_2\}^2 \div [\theta_0(1 - \theta_0)]. \qquad [*]$$

Write the second summand as $Q^2 = (n_1 n_2)\{a/n_1 - c/n_2\}^2 \div [N\theta_0(1 - \theta_0)]$. It is the negative exponent in the asymptotic (normal) density for the difference in sample "means." That is, asymptotically, a/n_1 and c/n_2 are independently a bivariate normal pair with a/n_1 normal $\mathbf{N}(\theta_0, \theta_0(1 - \theta_0) \div n_1)$, and c/n_2 normal $\mathbf{N}(\theta_0, \theta_0(1 - \theta_0) \div n_2)$. Hence, the quantity $(a/n_1 - c/n_2)$ is normally distributed $\mathbf{N}(0, \theta_0(1 - \theta_0)N \div n_1 n_2)$. Thus, Q^2 provides the "exact," probabilistic discrepancy for a significance test of the hypothesis that the coins are equally biased, using the sample difference in means as the test statistic. By Fisher's assumption that the marginal totals are irrelevant, this statistic exhausts the data.[6]

The "nuisance" parameter (θ_0) appears in Q^2 as part of the variance term; however, it may be removed by "Studentizing" the unknown Normal variance. Here is where the choice of estimate for θ_0 plays a role in the decomposition of χ^2. The left hand summand in [*] is easily recognized as $[\theta_0 - (a + c)/N]^2 \, I_{\theta_0}[N]$, where $I_{\theta_0}[N]$ denotes the Fisher Information about a binomial parameter contained in a sample of size N. Upon adopting an estimate T for θ_0, we see that the left hand summand in [*] becomes: $[T - (a + c)/N]^2 \, N \div T(1 - T)$. This term is positive for all estimates, except when T is the m.l.e for θ_0, that is, except when $T = (a + c) \div N$. Thus, the decomposition [*] of χ^2 provides a way of distinguishing among (1st order) efficient estimators. In short, the left hand term, $[T - (a + c)/N]^2 \, N \div T(1 - T)$, indicates the excess χ^2 discrepancy that results from using an estimator other than the m.l.e. For example, though total χ^2 is reduced by using the minimum χ^2 estimate compared with the m.l.e for θ_0, the minimum χ^2

estimate yields a test of significance with a distorted discrepancy in comparison with the "exact" (asymptotic) test. In fact, Fisher uses a similar decomposition of χ^2 in his discussion of rival (1st order) efficient estimates (SMfRW, §57, especially Figure 12).[7] Thus, we have arrived at one of Fisher's arguments for the supremacy of maximum likelihood estimation, based on Information, understood as a response to the challenge of data reduction without loss of relevant information.

2. INFORMATION AND THE DESIGN OF EXPERIMENTS

2.1. *Control and Precision*

Allow me to begin the second part of my talk with a typical "horror" story about the frustration that statisticians experience when they are called on to consult in "data analysis." A Professor and Graduate student from the Agriculture Division of a noted institution appear at our statistician's door with reams of data from an experiment concerning the (multi-attribute) yield of two varieties of corn. Their data have been carefully collected and recorded. But our satistician is at his wit's end. Alas, test variety 1 of corn was planted on field A, test variety 2 of corn was planted on field B, and the two fields are not alike! If only the researchers had consulted on the design first. With a few experimental controls their results could have been more revealing even at half the sample sizes!

Fisher [1962] expresses the same problem this way.

> When, a little more than 25 years ago, I first attempted a systematic exposition of the subject, known as the Design of Experiments, it is no very grave confession to avow that I did not fully understand the position among the statistical sciences of this new discipline. My approach at that time was frankly a technological one. As a statistician I had often set myself the task of analyzing experimental data, and was much concerned with those improvements in statistical methods which promised to make such analysis, more thorough and more comprehensive. Technically, I could see that some methods were superior to others in the concrete sense of extracting from the data more 'information' on the subject under enquiry, and therefore of leading to estimates of higher precision, and to tests of significance of greater sensitivity. And so it was in this atmosphere, borne in upon me that very often, when the most elaborate statistical refinements possible could increase the precision by only a few per cent, yet a different design involving little, or no additional experimental labour, might increase the precision two-fold, or five-fold or even more, and could often supply information in addition on relevant supplementary questions on which the original design was completely uninformative.

It was thus clear at an stage that there were quantitatively large technological gains to be obtained through the deliberate study of Experimental Design, and that these gains were to be harvested by making the plan of experimentation and observation logically coherent with the aims of the experiment, or, in other words with the kind of inference about the real world, which it was hoped that the experimental results would permit.

Fisher's thesis here is that when an experiment is improved, for example, by introducing better controls or other design considerations, the resulting data contain more Fisher-Information.

ILLUSTRATION. The research goal is to investigate the difference δ between two experimental treatments. There are two designs to choose from.

Experiment 1 — Arrange the field trials so that one treatment yields observations x_i $(i = 1, \ldots, n)$ which are i.i.d. normal $N(\mu, \sigma^2)$ and the second treatment yields observations y_i $(i = 1, \ldots, n)$ which are i.i.d. normal $N(\mu + \delta, \sigma^2)$. Suppose all three parameters are unknown, but that δ is the sole parameter of interest. This design might arise with random assignments to $2n$ plots in a given field. Half of the plots are randomly allocated to the first treatment and the remaining half are used for the second treatment.

Experiment 2 — Arrange the field trials so that $(x_i, y_i\ (i = 1, \ldots, n))$ are n-matched pairs in the field, producing data from a bivariate normal population with unknown correlation ρ. Thus, as before, the data x_i $(i = 1, \ldots, n)$ are i.i.d. normal $N(\mu, \sigma^2)$ and they y_i $(i = 1, \ldots, n)$ are i.i.d. normal $N(\mu + \delta, \sigma^2)$, but the (x_i, y_i) pairs have correlation ρ. This design might be implemented by blocking the $2n$ plots into n pairs and randomly assigning one plot/block to each of the two test groups.

Which is the better experiment? Let us apply Fisher-Information to the resulting test estimators with which we shall conduct our inferences about δ. With design 1 (random assignment) there is a simple "Student's" t-test, based on the sample difference $(\bar{Y} - \bar{X})$, and a "pooled" variance estimate having $2(n - 1)$ degrees of freedom. In the second design (randomized blocks), there is a simple t-test based on the sequence of differences $z_i = y_i - x_i$ $(i = 1, \ldots, n)$. Specifically, use a t-test with the sample average difference, $\bar{Z}$, and the associated variance estimate having $(n - 1)$ degrees of freedom.

Which design yields the test-statistic with the greater Fisher-Information about δ? The answer depends solely on the sample size n and the strength of the "matching," ρ. The table, below, lists the critical matching strengths (ρ) for which the second design has more Information about δ.

TABLE OF CRITICAL VALUES FOR ρ

n (number of paired observations)	ρ
2	0.167
3	0.160
4	0.143
5	0.127
10	0.0790
15	0.0569
26	0.0351
50	0.0188
250	0.0049
500	0.0020

Thus, the emphasis on match-pairs — a design concern — is justified by the fact that even with moderate samples sizes there is more Information about δ in the second design than in the first. Better to increase the precision by blocking (and use an estimate with n-1 degrees of freedom) rather than to double the degrees of freedom in a fully random allocation. Similar analysis indicates when Latin Squares, or other forms of restricted designs, yield estimates with greater Information about the parameters of interest. Information affords a measure of the efficacy in contemplated experimental controls.

2.2. *The Problem of Randomization*

Early in *The Design of Experiments* (§9—10) Fisher argues that randomization is a sine qua non of sound experimental practice. His now infamous pedagogical example, "The Lady Tasting Tea," offers three methodological lessons. We are to test the hypothesis that a certain lady cannot distinguish between tea made first by adding the milk as opposed to tea made by adding milk second.

The design calls for presenting her with 8 cups of tea with milk, 4

prepared each way, and counting how many the lady correctly identifies. According to Fisher's reasoning, by rigorously randomizing both the division of the cups (between the two treatments) and the order in which they are presented to the Lady, we achieve three goals:

(1) The random order of presentation is supposed to insure against the lady doing well on the test merely by anticipating the experimenter if, on an alternative design, the order of cups is decided by the experimenter. The randomization, then, is to prevent the experiment from becoming a game where the subject tries to outfox the experimenter.

(2) By randomizing the treatment allocation, the design is thought to insure against an unfortunate confounding of treatment with uncontrolled factors, which factors might be the actual cause of the lady's responses. For example, if the lady reacts to unobserved differences in the cups themselves, rather than to the tea mixtures, randomization establishes there is only a 1 in 70 chance the tea-milk combinations will align with cups so that she correctly identifies all 8.

(3) Randomizing the design, argues Fisher, provides a sound statistical basis for the resulting test of significance. That is, the randomization justifies the conclusion that, under the null hypothesis, there is a probability of 1/70 that all 8 cups are correctly identified, etc.

There is, however, a serious difficulty incorporating these arguments into the theory I am attributing to Fisher. That theory attempts to unify design with analysis, to use Information as the link between experimental design and statistical estimation. The problem centers on the role of ancillary data. (Recall, statistic T is ancillary for θ if it is probabilistically independent of the parameter of interest, $p(T|\theta) = p(T)$). In Fisher's theory of estimation, an ancillary statistic contains no relevant information about the parameter of interest. (The same conclusion follows according to Bayesian or Likelihood principles.) In numerous places Fisher takes pains to argue that, by conditioning on an ancillary statistic one creates estimates with greater Information.[8] Also, as in the illustration of the 2×2 table, conditioning on so-called "ancillary" data (the margin totals) may create an "exact" significance test for some composite null hypothesis.

The difficulty is simple to state: Randomization in design introduces ancillary data. That is, the outcomes of the randomization is ancillary to the hypothesis tests. Unfortunately, each of Fisher's three reasons for randomized design is based on probabilities which fail once the ancillary data of the randomization are given. That is, Fisher's argu-

ments for randomization in design are valid using pretrial expectations; but they are unsupported post-trial, given the (ancillary) randomized outcome. There is no question that Fisher opposed randomization in analysis. He used the ancillarity principle to refute randomized statistical tests.[9] Nonetheless, his support for randomization in design remained undaunted. Even mild opposition from his longtime ally, Gosset ("Student," 1936), earned only (undeserved) scorn (Fisher, 1936a, b).[10] Among the several enigmas we owe to Fisher, coming to a proper understanding of the role randomization plays in sound experimental design is an ongoing activity. (See, for example, Rubin 1978.)

3. CONCLUSIONS

The position proposed here is that Fisher's contributions to the theory of experimental design are closely tied to his theory of estimation, with Information serving as the link. Estimation is concerned with data-reduction — where good estimators give summaries of evidence that preserve the relevant evidence as that is measured by Fisher-Information. Improvements in a design may be gauged by the increase in (Fisher-) Information of the resulting experimental data. Thus, there is a unified approach to statistical design and analysis. This approach serves, also, to explain the widely received view by statisticians that their role is not limited to post-trial consultation in data analysis. Statistics has its place in the planning of high quality experiments. In that sense, for an experimenter, it makes statistical sense to look before you leap!

The reconstruction I offer has difficulty, however, providing a rationale for the common methodological practice of randomized experimental design. The problem is that, in estimation Fisher's theory advocates conditioning on ancillary data. But randomization yields ancillary data. Then familiar (Fisherian) arguments fail when the same ancillary principle is applied to the outcome of the randomization. It is reassuring, at least, to know that the debates about randomization persist more than 55 years after Fisher's methodological innovation. (See, for example, the papers by I. Levi, D. Lindley, and P. Suppes in PSA-1982.)

Carnegie Mellon University

NOTES

[1] In SMfRW it is chapter 9, making up 40 pages out of 360. In DoE it is chapter 11, making up 35 out of 245 pages. And in SM&SI it is chapter 6, 35 of 180 pages.

[2] Fisher's earlier attempts to define consistency asymptotically, with increasing sample size, failed as they placed no restriction on how the estimator behaved in small samples. The version of F-consistency summarized here applies equally well to samples of arbitrary sizes, as Fisher notes (SM&SI, pp. 150—151).

[3] I find this assumption troubling. It seems plausible that the sample size is chosen in accord with the investigator's pretrial beliefs about the informativeness of the resulting data. But, except in rare circumstances, this judgment depends then upon the investigator's "prior" opinions about the parameter of interest. Hence, as a reader of the published data, I conclude that sample size is a function of the unknown parameter through the experimenter's "prior" for that parameter. That is, as a reader of the published data, I cannot take N to be ancillary without defaming the experimenter!

[4] See Savage (1976) for definitive rebuttals to these inaccuracies.

[5] See Fisher's statements in (SM&SI, §4.4) and (SMfRS, §21.02). The question is pursued by G. Barnard in three papers spanning the years 1946—1949.

[6] That is, given the sample sizes n_1 and n_2, the two quantities $(a/n_1 - c/n_2)$ and $(a/n_1 + c/n_2)$ are equivalent to the full data (a, b, c, d). Fisher's stipulation that the analysis be conducted for fixed lower margins amounts to a further data reduction to the test quantity, $(a/n_1 - c/n_2)$.

[7] The term $[T - (a + c)/N]^2 N \div T(1 - T)$ differs by a factor of N from C. R. Rao's measure of 2nd order efficiency. First order efficiency concerns the asymptotic ratio of information retained in an estimate. Second order efficiency concerns the difference between information retained and information available. Obviously, this limiting ratio may be 1 though the limiting difference is not 0. For example, minimum χ^2 estimation *is not*, though the m.l.e *is* 2nd order efficient in the 2×2 table. See Rao (1963) and Ghosh and Subramanyam (1974) for the key results.

[8] See especially Fisher's discussion of what he calls "The Problem of the Nile" (SM&SI §6.9).

[9] Fisher's opposition to a randomized solution of the Behrens-Fisher problem is found in SM&SI, §4.7. Additional discussion of this example may be found in Kadane & Seidenfeld (1990).

[10] In fairness, I believe Fisher allowed Gosset the last word in this exchange. Fisher left "Student's" (1937) final publication unanswered. That gesture, rare for Fisher, signifies his lasting respect for Gosset's contributions.

REFERENCES

Barnard, G. (1946), 'Significance Tests for 2×2 Tables', *Biometrika* **34**, 123—138.

Barnard, G. (1947), 'The Meaning of a Significance Level', *Biometrika* **34**, 179—182.

Barnard, G. (1949), 'Statistical Inference', *J. Royal Stat. Soc.* **B 11**, 115—140.

Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton: Princeton Univ. Press.

Fisher, R. A. (1922), 'On the Mathematical Foundations of Theoretical Statistics', *Phil. Trans. Roy. Soc. London* **A 22**, 309—368.
Fisher, R. A. (1925a), 'Theory of Statistical Estimation', *Proc. Cambridge Phil. Soc.* **22**, 700—725.
Fisher, R. A. (1925b), *Statistical Methods for Research Workers* (14th ed., 1973), New York: Hafner Press.
Fisher, R. A. (1934), 'Two New Properties of Mathematical Likelihood', *Proc. Roy. Soc. London* **A 144**, 285—307.
Fisher, R. A. (1935), *The Design of Experiments* (8th ed., 1966), New York: Hafner Press.
Fisher, R. A. (1936a), 'A Test of the Supposed Precision of Systematic Arrangements', *Annals of Eugenics* **7**, 189—193.
Fisher, R. A. (1936b), 'The Half-Drill Strip System Agricultural Experiments', *Nature* **138**, 1101.
Fisher, R. A. (1956), *Statistical Methods and Scientific Inference* (3rd ed., 1973), New York: Hafner Press.
Fisher, R. A. (1962), 'The Place of the Design of Experiments in the Logic of Scientific Inference', *Colloques Internationaux du Centre National de la Recherche Scientifique (Paris)* **110**, 13—19.
Ghosh, J. K. and Subramanyam, K. (1974), 'Second Order Efficiency of Maximum Likelihood Estimators', *Sankhya* **36**, 325—358.
Kadane, J. B. and Seidenfeld, T. (1990), 'Randomization in a Bayesian Perspective', *J. Stat. Planning and Inference* **25**, 329—345.
Levi, I. (1983), 'Direct Inference and Randomization', in P. Asquith and T. Nickles (eds.), *PSA-1982*, vol. 2. Ann Arbor: Edwards Brothers, 447—463.
Lindley, D. (1983), 'The Role of Randomization in Inference', in P. Asquith and T. Nickles (eds.), *PSA-1982*, vol. 2. Ann Arbor: Edwards Brothers, 431—446.
Rao, C. R. (1963), 'Criteria of Estimation in Large Samples', *Sankhya* **25**, 189—206.
Rubin, D. (1978), 'Bayesian Inference for Causal Effects: The Role of Randomization', *Ann. Stat.* **6**, 34—58.
Savage, L. J. (1976), 'On Rereading R. A. Fisher', *Ann. Stat.* **4**, 441—500.
"Student" (1936), 'Co-operation in Large-Scale Experiments', opening remarks by W. S. Gosset, Supplement to *Roy. Stat. Soc.* **3**, 115.
"Student" (1937), 'Random and Balanced Arrangements', *Biometrika* **29**, 363—379.
Suppes, P. (1983), 'Arguments for Randomizing', in P. Asquith and T. Nickles (eds.), *PSA-1982*, vol. 2. Ann Arbor: Edwards Brothers, 464—475.