Conditioning, Likelihood, and Coherence: A Review of Some Foundational Concepts

James Robins & Larry Wasserman

309–368.

——— (1934), "Two New Properties of Mathematical Likelihood," *Proceedings of the Royal Society*, Ser. A, 144, 285–307.

——— (1956), *Statistical Methods and Scientific Inference*, Edinburgh: Oliver and Boyd.

Fraser, D. A. S., Reid, N., and Wu, J. (1999), "A Simple General Formula for Tail Probabilities for Frequentist and Bayesian Inference," *Biometrika*, 86, 249–264.

Geyer, C. (1999), "Likelihood Inference for Spatial Point Processes," in *Stochastic Geometry: Likelihood and Computation*, eds. O. E. Barndorff-Nielsen, W. S. Kendall, and M. N. M. van Lieshout, London: Chapman and Hall/CRC, pp. 79–140.

Hinkley, D. V. (1980), "Likelihood as Approximate Pivotal," *Biometrika*, 67, 287–292.

Jeffreys, H. (1961), *Theory of Probability*, Oxford, U.K.: Oxford University Press.

Kalbfleisch, J. G. (1985), *Probability and Statistics*, Vol. II (2nd ed.), New York: Springer-Verlag.

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Kass, R. E., and Wasserman, L. (1996), "Formal Rules for Selecting Prior Distributions: A Review and Annotated Bibliography," *Journal of the American Statistical Association*, 91, 1343–1370.

Li, B., and McCullagh, P. (1994), "Potential Functions and Conservative Estimating Functions," *The Annals of Statistics*, 22, 340–356.

Lindley, D. V. (1958), "Fiducial Distributions and Bayes's Theorem," *Journal of the Royal Statistical Society*, Ser. B, 20, 102–107.

Lugannani, R., and Rice, S. (1980), "Saddlepoint Approximation for the Distribution of the Sum of Independent Random Variables," *Advances in Applications and Probabilities*, 12, 475–490.

McCullagh, P. (1983), "Quasi-Likelihood Functions," *Annals of Statistics*, 11, 59–67.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

Mukerjee, R., and Reid, N. (1999), "On Confidence Intervals Associated With the Usual and Adjusted Likelihoods," *Journal of the Royal Statistical Society*, Ser. B, 61, 945–953.

Murphy, S. A., and van der Vaart, A. (1997), "Semiparametric Likelihood Ratio Inference," *The Annals of Statistics*, 25, 1471–1509.

Owen, A. (1988), "Empirical Likelihood Ratio Confidence Intervals for a Single Functional," *Biometrika*, 75, 237–249.

Redelmeier, D., and Tibshirani, R. J. (1997a), "Association Between Cellular Phones and Car Collisions," *New England Journal of Medicine*, February 12, 1997, 1–5.

——— (1997b), "Is Using a Cell Phone Like Driving Drunk?," *Chance*, 10, 5–9.

——— (1997c), "Cellular Telephones and Motor Vehicle Collisions: Some Variations on Matched Pairs Analysis," *Canadian Journal of Statistics*, 25, 581–593.

Reid, N. (1995), "The Roles of Conditioning in Inference," *Statistics in Science*, 10, 138–157.

——— (1996), "Likelihood and Higher-Order Approximations to Tail Areas: A Review and Annotated Bibliography," *Canadian Journal of Statistics*, 24, 141–166.

Royall, R. (1997), *Statistical Evidence*, London: Chapman and Hall.

Severini, T. A. (1998), "An Approximation to the Modified Profile Likelihood Function," *Biometrika*, 82, 1–23.

Skovgaard, I. M. (1990), "On the Density of Minimum Contrast Estimators," *The Annals of Statistics*, 18, 779–789.

——— (1996), "An Explicit Large-Deviation Approximation to One-Parameter Tests," *Bernoulli*, 2, 145–165.

Welch, B. L., and Peers, H. W. (1963), "On Formulae for Confidence Points Based on Integrals of Weighted Likelihoods," *Journal of the Royal Statistical Society*, Ser. B, 25, 318–329.

Conditioning, Likelihood, and Coherence: A Review of Some Foundational Concepts

James ROBINS and Larry WASSERMAN

1. INTRODUCTION

Statistics is intertwined with science and mathematics but is a subset of neither. The "foundations of statistics" is the set of concepts that makes statistics a distinct field. For example, arguments for and against conditioning on ancillaries are purely statistical in nature; mathematics and probability do not inform us of the virtues of conditioning, but only on how to do so rigorously. One might say that foundations is the study of the fundamental conceptual principles that underlie statistical methodology. Examples of foundational concepts include ancillarity, coherence, conditioning, decision theory, the likelihood principle, and the weak and strong repeated-sampling principles. A nice discussion of many of these topics was given by Cox and Hinkley (1974).

There is no universal agreement on which principles are "right" or which should take precedence over others. Indeed, the study of foundations includes much debate and controversy. An example, which we discuss in Section 2, is the likelihood principle, which asserts that two experiments that yield proportional likelihood functions should yield identical inferences. According to Birnbaum's theorem, the likelihood principle follows logically from two other principles: the conditionality principle and the sufficiency principle. To many statisticians, both conditionality and sufficiency seem compelling yet the likelihood principle does not. The mathematical content of Birnbaum's theorem is not in question. Rather, the question is whether conditionality and sufficiency should be elevated to the status of "principles" just because they seem compelling in simple examples. This is but one of many examples of the type of debate that pervades the study of foundations.

This vignette is a selective review of some of these key foundational concepts. We make no attempt to be complete

in our coverage of topics. In Section 2 we discuss the likelihood function, the likelihood principle, the conditionality principle, and the sufficiency principle. In Section 3 we briefly review conditional inference. In Section 4 we discuss Bayesian inference and coherence. In Section 5 we provide a brief look at some newer foundational work that suggests that in complex problems, the conditionality principle, the likelihood principle, and coherence arguments may be less compelling than in the low-dimensional problems, where they are usually discussed.

2. THE LIKELIHOOD FUNCTION AND THE LIKELIHOOD PRINCIPLE

Consider a random variable Y and a model $\mathcal{M} = \{p_\theta(\cdot); \theta \in \Theta\}$ for the distribution of Y. Here each $p_\theta(\cdot)$ is a density for Y, θ is an unknown parameter, and the parameter space Θ is a subset of $\mathcal{R}^k$. Assume that we have n iid replicates of Y denoted by $Y^n = (Y_1, \ldots, Y_n)$, generated from p_{θ_0}, where $\theta_0 \in \Theta$ denotes the true value of the unknown parameter θ. Fisher (1921, 1925, 1934) defined the likelihood function as

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} p_\theta(Y_i).$$

Of course, the likelihood function appeared implicitly much earlier in the work of Bayes and Laplace, who used what we now call Bayesian inference to solve statistical problems. But it was Fisher who first emphasized the essential role of the likelihood function itself in inference. (See Aldrich 1997, Edwards 1997, and Fienberg 1997, for discussions on the history of likelihood.)

From a Bayesian perspective, inferences are based on the posterior $p(\theta|Y^n) \propto \mathcal{L}_n(\theta)\pi(\theta)$, where $\pi(\theta)$ is a prior distribution for θ. In the Bayesian framework, the data enter the inferences only through the likelihood function.

From a frequentist perspective, the likelihood is a source of point and interval estimators. For example, the maximum likelihood estimator (MLE) $\hat{\theta}_n$—the point at which $\mathcal{L}_n(\theta)$ takes its maximum—is known, under weak conditions, to be asymptotically normal with the smallest possible asymptotic variance. The set $C_n(c) = \{\theta; \mathcal{L}_n(\theta)/\mathcal{L}_n(\hat{\theta}_n) \leq c\}$ gives the asymptotically shortest, $1 - \alpha$ confidence set if c is chosen appropriately, if θ is scalar.

The likelihood function is central to many statistical analyses. More controversial is the role of the likelihood principle (LP). The LP says that when two experiments yield proportional likelihoods they should yield identical inferences. This principle of inference is accepted as a guiding principle by some statisticians but is considered unreasonable by others. Most forms of frequentist inference, such as significance testing and confidence intervals, violate the LP and so are ruled out if one wishes to follow the LP. For example, consider the likelihood-based confidence interval $C_n(c)$ defined earlier. Suppose that two different experiments yielded proportional likelihoods. Then the form of the interval $C_n(c)$ would be the same in the two experiments, but the coverage probability ascribed to $C_n(c)$ could be different, thus violating the LP.

A famous example that illustrates the likelihood principle involves binomial versus negative binomial sampling. Suppose that we want to estimate the probability θ of "heads" for a coin. In the binomial experiment, we toss the coin n times and observe the number of "heads" Y. Here n is fixed and Y is random. In the negative binomial experiment, we observe the number of tosses N required to obtain y "heads." Here N is random and y is fixed. Suppose that we observe 3 heads in 5 tosses. Under binomial sampling, these data yield the likelihood function

$$\mathcal{L}_1(\theta) = \binom{5}{3} \theta^3 (1 - \theta)^2.$$

Under negative binomial sampling, the likelihood function is

$$\mathcal{L}_2(\theta) = \binom{4}{2} \theta^3 (1 - \theta)^2.$$

As functions of θ, these likelihood functions are proportional, so any method that obeys the likelihood principle should yield identical inferences about θ regardless of whether the data were obtained by binomial or negative binomial sampling. Frequentist confidence intervals violate the likelihood principle, because the coverage of an interval is evaluated under hypothetical repetitions of the experiment. The set of possible outcomes in these hypothetical repetitions will differ, depending on whether binomial or negative binomial sampling is used. An interesting discussion of this problem was provided by Lindley and Phillips (1976).

Birnbaum (1962) showed that the likelihood principle follows logically from two other principles: the conditionality principle (CP) and the sufficiency principle (SP). A clear exposition of the details of Birnbaum's theorem and its implications was given by Berger and Wolpert (1984). This and the next section draw heavily from that monograph (see also Cox and Hinkley 1974).

The conditionality principle (Cox 1958) says that if we decide which of two experiments to do by the flip of a fair coin, then the final inference should be the same as if the experiment had been chosen without flipping the coin. More formally, the (weak) form of CP can be described as follows. Suppose that we consider two experiments, E_1 and E_2, for the same parameter θ. We flip a fair coin and perform E_1 if the coin is "heads" and perform E_2 if the coin is "tails." This is called a "mixed experiment." CP asserts that if we obtain "heads" and perform E_1, then our inferences should be identical to the inferences that we would make if E_1 were performed without first flipping the coin (and similarly for E_2). The outcome of the coin flip is an example of an ancillary statistic; that is, a statistic whose distribution does not depend on the unknown parameter. The sufficiency principle says that two outcomes of an experiment that yield the same value of a sufficient statistic should yield identical inferences. Many statisticians find CP and SP quite appealing, yet they do not find LP appealing, despite the fact the LP follows logically from these two

principles. In fact, Evans, Fraser, and Monette (1986) since showed that LP follows from a slightly stronger version of CP alone.

The LP and CP have many supporters and detractors. Often, statisticians who find Bayesian methods appealing favor the LP. A non-Bayesian approach that obeys the LP was given by Edwards (1972). Those who prefer frequentist methods, as developed by Neyman, Wald, and others, find LP less compelling. Let us add our own point of view. The CP seems compelling in simple examples, provided that the experiment performed gives no additional information about the parameter beyond that contained in the data. In Section 5 we show that the CP is less compelling in high-dimensional models.

3. CONDITIONING, ANCILLARITY, AND RELEVANT SUBSETS

In the preceding section, when we discussed the conditionality principle, it seemed natural that inferences in the mixed experiment should be made conditionally on the value of the coin flip. Such inferences would then obey CP. This raises a more general question: Should inferences be carried out conditionally on an appropriate statistic? Inferences made conditional on some statistic go under the rubric of "conditional inference." Conditional inference can be traced back to Fisher (1956), Cox (1958), and others. The appeal of conditioning is evident from the following simple example (example 1 of Berger and Wolpert 1984). We observe two iid random variables Y_1 and Y_2, where $P_\theta(Y_i = \theta - 1) = P_\theta(Y_i = \theta + 1) = 1/2, i = 1, 2$. Here θ is an unknown real number. Let $C = \{(Y_1 + Y_2)/2\}$ if $Y_1 \neq Y_2$ and $C = \{Y_1 - 1\}$ otherwise. C is a 75% confidence set, although, unlike the typical confidence set, it contains only a single point. Thus $P_\theta(\theta \in C) = .75$ for all θ. But when $Y_1 \neq Y_2$, we are certain that $\theta \in C$ and when $Y_1 = Y_2, C$ contains θ 50% of the time; that is, $\Pr(\theta \in C|Y_1 = Y_2) = 1/2$. This suggests partitioning the sample space into $\{B, B^c\}$, where $B = \{Y_1 \neq Y_2\}$, and then reporting different inferences depending on whether B occurs or does not occur. This is equivalent to reporting inferences conditional on the statistic $S = |Y_1 - Y_2|$. The example is meant to suggest that inferences will be more intuitively plausible if they are performed conditional on an appropriate statistic.

If we accept that inferences should sometimes be conditional on something, then that raises the question of what we should condition on. The most common choices to condition on are ancillary statistics and relevant subsets. Both concepts were discussed by Fisher (1956).

An ancillary statistic is a statistic whose distribution does not depend on θ. In the foregoing example, $S = |Y_1 - Y_2|$ is ancillary, and it seems quite reasonable to report a confidence of 1 when $S = 1$ and a confidence of .50 when $S = 0$. Specifically, we report $P_\theta(\theta \in C|S = 1) = 1$ and $P_\theta(\theta \in C|S = 0) = .5$. The coin flip in the mixed experiment described in Section 2 is another example of an ancillary statistic.

Consider a $1 - \alpha$ confidence set $C(Y)$; that is, $P_\theta(\theta \in C(Y)) = 1 - \alpha$ for all θ. We say that B is a relevant subset if there is some $\varepsilon > 0$ such that either $P_\theta(\theta \in C(Y)|Y \in B) \leq (1 - \alpha) - \varepsilon$ for all θ or $P_\theta(\theta \in C(Y)|Y \in B) \geq (1 - \alpha) + \varepsilon$ for all θ. If a relevant subset exists, then it seems tempting to report different inferences depending on whether $Y \in B$ or $Y \notin B$. Buehler and Fedderson (1963) and Brown (1967) showed that there exists a relevant subset even for the familiar Student t intervals for the mean of a normal. Thus the existence of relevant subsets is far from pathological.

Bayesian inference is an extreme form of conditional inference, because the posterior conditions on the data itself, as opposed to a conditioning on some statistic. There have been attempts to build formal, non-Bayesian theories of conditional inference. The best-known attempt is probably that of Kiefer (1977). Other contributions have come from Brown (1967, 1978), Buehler (1959, 1976), Casella (1987), Cox (1958, 1980), Fraser (1977), and Robinson (1976, 1979). The main idea is to partition the sample space as $\mathcal{Y} = \cup_s \mathcal{Y}_s$ and report conditional confidence $P_\theta(\theta \in C(Y)|Y \in \mathcal{Y}_s)$ when $Y \in \mathcal{Y}_s$.

Conditional inference is appealing because it seems to solve the apparently counterintuitive results like those in the simple example presented at the beginning of this section. Nevertheless, in many models there is no known ancillary or relevant subset on which to condition, or there may be many ancillaries, in which case it is not clear on which to condition. One way to extend the applicability of conditional inference when there is no exact ancillary is to use approximate conditional inference, in which one conditions on a statistic that is asymptotically ancillary (see, e.g., Amari 1982; Barndorff-Nielsen 1983; Cox 1988; Cox and Reid 1987; DiCiccio 1986; Efron and Hinkley 1978; Robins and Morgenstern 1987; Severini 1993; Sweeting 1992).

Brown (1990) and Foster and George (1996) gave examples in which an estimator is admissible conditional on an ancillary statistic but is unconditionally inadmissible. Such examples show explicitly that procedures that obey the CP can have poor unconditional properties. We discuss this point further in Section 5.

Another situation where conditioning has received attention is in the problem of estimating a common odds ratio ψ in a series of 2×2 tables. Specifically, we observe independent binomial random variables $X_{0k} \sim \text{bin}(n_{0k}, p_{0k})$ and $X_{1k} \sim \text{bin}(n_{1k}, p_{1k})$ $k = 1, \ldots, K$, where $p_{1k} = p_{0k}/(p_{0k} + \psi(1 - p_{0k}))$. The likelihood is $L_1(\psi)L_2(\psi, \mathbf{p_0})$, where

$$L_1(\psi) = f(\mathbf{X}_1|\mathbf{X}_+; \psi) = \prod_{k=1}^{K} f(X_{1k}|X_{+k}; \psi)$$

and

$$L_2(\psi, \mathbf{p_0}) = f(\mathbf{X}_+; \psi, \mathbf{p_0}) = \prod_{k=1}^{K} f(X_{+k}; \psi, p_{0k}),$$

where $\mathbf{p_0} = (p_{01}, \ldots, p_{0K}), \mathbf{X}_1 = (X_{11}, \ldots, X_{1k}), \mathbf{X}_+ = (X_{+1}, \ldots, X_{+K})$, and $X_{+k} = X_{0k} + X_{1k}$. Now the set of

row totals $\mathbf{X}_+$ is not S ancillary for ψ; that is, there is no global reparameterization (ψ, θ) such that ψ and θ are variation independent (i.e., the parameter space is a product space) and such that $f(\mathbf{X}_+; \psi, \theta) = f(\mathbf{X}_+; \theta)$. Thus the CP does not imply that inference for ψ should be performed conditional on $\mathbf{X}_+$. However, it has been argued that inference for ψ (conditional or unconditional) should be based on the conditional likelihood $L_1(\psi)$ if the marginal likelihood $L_2(\psi, \mathbf{p}_0)$ contains no independent information about ψ when $\mathbf{p}_0$ is unknown. The conditional MLE maximizing $L_1(\psi)$ is asymptotically efficient for ψ both in large-stratum asymptotics, in which $n_{0k} \to \infty$ and $n_{1k}/n_{0k} \to c_k$ for each k and in a sparse data asymptotics, in which $K \to \infty, n_{1k}$ and n_{0k} are bounded and the p_{0k} are drawn independently from a common distribution (Bickel, Klaassen, Ritov, and Wellner 1993; Lindsay 1980). Thus, asymptotically, $L_2(\psi, \mathbf{p}_0)$ contains no additional information about ψ. However, Sprott (1975) provided an interesting example to show that $L_2(\psi, \mathbf{p}_0)$ can contain some information about ψ. He considered the special case where $n_{1k} = n_{0k} = 1$. Suppose that only $\mathbf{X}_+$ is observed and $X_{+k} = 1$ for all k. Then he argued that for large K, the hypothesis $\psi = 1$ can be rejected, because for any p_{0k}, the probability that $X_{+k} = 1$ when $\psi = 1$ can never exceed $1/2$. Of course, the information about ψ contained in the marginal law of $\mathbf{X}_+$ is asymptotically negligible as $K \to \infty$ compared to that in $L_1(\psi)$, because the conditional MLE is efficient.

Finally, we should add that conditioning is used for other reasons as well; for example, in the construction of similar tests (Cox and Hinkley 1974, chap. 5; Lehmann 1986, chap. 4). A general discussion of conditional inference was provided by Lehmann (1986, chap. 10).

4. COHERENCE AND BAYESIAN INFERENCE

Some researchers have attempted to create a foundationally sound method of inference by stating axioms for inference and then characterizing all inferential methods that satisfy these axioms. Often, such axioms are called axioms of coherence, as they are meant to capture what a coherent (i.e., self-consistent) inference is. Usually, these arguments lead to conclusions of the form that inferences are coherent if and only if they are Bayesian. It may appear that this line of research has had a greater impact on statistical practice than conditioning arguments, because Bayesian methods have become increasingly popular in practice, whereas conditional inference has not. However, we believe the increasing use of Bayesian methods has more to do with their conceptual simplicity as well as advances in computing than with the coherence arguments. Still, these arguments do add interesting insight into inferential issues.

There are many versions of coherence arguments (see, e.g., de Finetti 1974, 1975; Freedman and Purves 1969; Heath and Sudderth 1978, 1989; Jeffreys 1961; Ramsey 1930; Regazzini 1987; Savage 1954). Here we describe the Heath–Sudderth approach. We begin with a model $\{P_\theta; \theta \in \Theta\}$, where each P_θ is a probability distribution for a random variable Y. An inference Q is a map that assigns a (possibly finitely additive) probability measure Q_y over Θ

to each outcome y. The function Q_y is regarded as a set of probabilities from which bets are made about θ after observing $Y = y$. An inference Q is called "coherent" if it is impossible to place a finite number of bets on subsets of Θ on observing $Y = y$, which have a strictly positive expected payoff. Heath and Sudderth proved that an inference Q is coherent if and only if it is a posterior for some, possibly finitely additive, prior π over Θ. Formally, Q is a posterior distribution for the prior π, if for every bounded, measurable function $\phi(\theta, y)$,

$$\int \int \phi(\theta, y) P_\theta(dy)\pi(d\theta) = \int \int \phi(\theta, y) Q_y(d\theta) m(dy).$$

Here, m is the marginal distribution for Y induced by the model and prior; that is, $\int g(y)m(dy) = \int \int g(y)P_\theta(dy)\pi(d\theta)$ for every bounded, measurable function $g(y)$.

The implication is that inferences are coherent if and only if they are Bayesian. Such results increase the appeal of Bayesian methods to many statisticians. Of course, the results are only as compelling as the axioms. Given the choice between a method that is coherent and a method that has, say, correct frequentist coverage, many statisticians would choose the latter. The issue is not mathematical in nature. The question is under which circumstances one finds coherence or correct coverage more important.

In some cases it is possible both to be coherent and to have good frequentist properties, in large samples. Indeed, in a finite-dimensional model, with appropriate regularity conditions, the following facts are known. Let $\hat{\theta}_n$ be the MLE, let $\bar{\theta}$ be the posterior mean, and let $Q(d\theta|Y^n)$ be the posterior based on n iid observations $Y^n = (Y_1, \ldots, Y_n)$. Then the following hold:

1. $\bar{\theta}_n - \hat{\theta} = O_P(n^{-1})$.
2. There exist regions C_n such that both $\int_{C_n} Q(d\theta|Y^n) = 1 - \alpha$, and C_n has frequentist coverage $1 - \alpha + O(n^{-1})$.
3. If N is any open, fixed Euclidean neighborhood of the true value θ_0, then $\int_N Q(d\theta|Y^n)$ tends to 1 almost surely.
4. The posterior concentrates around the true value θ_0 at rate $n^{-1/2}$; that is, $Q(\{\theta; |\theta - \theta_0| \geq a_n n^{-1/2}\}|Y^n) = o_P(1)$ for any sequence $a_n \to \infty$.

In words, (1) the MLE and posterior mean are asymptotically close, (2) Bayesian posterior intervals and confidence intervals agree asymptotically, (3) the posterior is consistent, and (4) the posterior converges at the same rate as the maximum likelihood estimate. Facts 1, 3, and 4 follow from standard asymptotic arguments (see, Schervish 1995). Fact 2 was shown by Welch and Peers (1963), for the more difficult one-sided case.

In infinite-dimensional models, the situation is less clear. Consistency is sometimes attainable and sometimes not (see, e.g., Barron 1988; Barron, Schervish, and Wasserman 1999; Diaconis and Freedman 1986, 1990, 1993, 1997; Doob 1949; Freedman 1963, 1965; Freedman and Diaconis 1983; Ghosal, Ghosh, and Ramamoorthi 1999; Schwartz 1960, 1965) for example. Similarly, good rates of convergence are sometimes possible (see Ghosal, Ghosh, and van der Vaart 1998; Shen and Wasserman 1998; Zhao 1993,

1998). The issue of matching posterior probability and frequentist coverage has received less attention. While some negative results were reported by Cox (1993) and Freedman (1999), this topic remains mostly unexplored territory.

5. A LOOK TO THE FUTURE: FOUNDATIONS IN INFINITE-DIMENSIONAL MODELS

For the most part, foundational thinking has been driven by intuition based on low-dimensional parametric models. But in modern statistical practice, it is routine to use high-dimensional or even infinite-dimensional (nonparametric or semiparametric) methods. There is some danger in extrapolating our intuition from finite-dimensional to infinite-dimensional models. Should we rethink foundations in light of these methods? Here we argue that the answer is "yes." We summarize an example that was discussed in great detail by Robins and Ritov (1997). To keep things brief and simple, we give a telegraphic version and omit the details; see the Robins and Ritov article for a full discussion. Although it will not be immediately obvious, the example stems from a real problem—the analysis of treatment effects in randomized trials. See also Robins, Rotnitzky, and Van der Laan (2000).

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be n iid copies of a random vector (X, Y), where X is continuous taking values in the k-dimensional unit cube $\mathcal{X} = (0, 1)^k$ and Y given $X = x$ is normal with mean $\theta_0(x)$ and variance 1. The conditional mean function $\theta_0: (0, 1)^k \to \mathcal{R}$ is assumed to be continuous and to satisfy $\sup_{x \in (0,1)^k} |\theta(x)| \le M$ for some known positive constant M. Let Θ denote all such functions. The density $f_0(x)$ of X is assumed to belong to the set of densities

$$\mathcal{F}_X = \{f; c < f(x) < 1/c \text{ for } x \in \mathcal{X}\},$$

where $c \in (0, 1)$ is a fixed constant. Our goal is to estimate the parameter $\psi_0 = \int_{\mathcal{X}} \theta_0(x)\, dx$. A pair (θ, f) completely determines a law of (X, Y). The likelihood function is

$$\mathcal{L}(\theta, f) = \mathcal{L}_1(\theta)\mathcal{L}_2(f)$$

$$= \left\{\prod_{i=1}^{n} \phi(Y_i - \theta(X_i))\right\} \left\{\prod_{i=1}^{n} f(X_i)\right\},$$

where $\phi(\cdot)$ denotes the standard normal density. Notice that in this likelihood, the parameters θ and f are functions. The model is infinite dimensional because the set Θ cannot be put into a smooth, one-to-one correspondence with a finite-dimensional Euclidean space. Let $\mathbf{X} = \{X_i; i = 1, \ldots, n\}$ denote the observed X_i's. When f_0 is known, $\mathbf{X}$ is ancillary. When f_0 is unknown, $\mathbf{X}$ is still ancillary but in a slightly different sense. Technically, $\mathbf{X}$ is S ancillary for ψ, because the conditional likelihood given $\mathbf{X}$ is a function of θ alone and the marginal likelihood of $\mathbf{X}$ is a function of f alone, θ and f are variation independent (i.e., the parameter space is a product space), and ψ is a function of θ only (Barndorff Nielsen 1978; Cox and Hinkley 1974).

Now we shall see that whether or not we know the distribution f_0 of the ancillary $\mathbf{X}$ has drastic implications for inference, in contrast to the usual intuition about an-

cillarity. When f_0 is unknown, Robins and Ritov (1997) showed that no uniformly consistent estimator of ψ_0 exists. See also Ritov and Bickel (1992). But when f_0 is known, there do exist uniformly consistent estimators of ψ_0. In fact, there exist estimators that are $\sqrt{n}$-consistent uniformly over all $\theta \times f \in \Theta \times \mathcal{F}_X$. For example, define the random variable $V = Y/f_0(X)$. Then $\bar{V} = n^{-1}\sum_{i=1}^{n} V_i = n^{-1}\sum_{i=1}^{n} Y_i/f_0(X_i)$ is uniformly $\sqrt{n}$-consistent. (Uniformity is important because it links asymptotic behavior to finite-sample behavior. This is especially important in high-dimensional examples; i.e., when k is large.)

This result has implications for many common inferential methods. In particular, standard likelihood-based and Bayesian estimators methods will fail to be uniformly consistent. To see this, note that maximum likelihood inference, profile likelihood inference, and Bayesian inference with independent priors on θ and f, all share the following property: They provide the same inferences for ψ whatever the known $f_0 \in \mathcal{F}_X$ that generated the data. We call such methods *strict factorization-based* (SFB) methods. Indeed, in the model with f_0 known, any inference method that satisfies the likelihood principle is SFB. Robins and Ritov (1997) showed that no SFB estimator can be consistent for ψ_0 uniformly over $(\theta_0, f_0) \in \Theta \times \mathcal{F}_X$.

The deficiencies in SFB methods extend to interval estimation. Any interval estimator that is not a function of f_0 will not be "valid." By valid, we mean that under all $(\theta_0, f_0) \in \Theta \times \mathcal{F}_X$ the coverage is at least $(1 - \alpha)$ at each sample size n and the expected length goes to 0 with increasing sample size. There are valid interval estimators for ψ, but these depend on f_0 and hence are not SFB, and they violate LP. An example of such an interval estimator is $\bar{V} \pm dn^{-1/2}$, where

$$d^2 = \frac{M^2 + 1}{1 - \alpha} \int_{\mathcal{X}} \frac{dx}{f_0(x)}.$$

That this has coverage exceeding $1-\alpha$ follows from Chebyshev's inequality. Note that this interval is not only valid, but its length shrinks at rate $n^{-1/2}$. However, even with f_0 known, there is no interval estimator that has expected length tending to 0, with conditional coverage at least $1-\alpha$ given $\mathbf{X}$, on a set of $\mathbf{X}$ with f_0 probability 1, for all $\theta_0 \in \Theta$. Our example is connected to Godambe and Thompson's (1976) criticism of likelihood-based inference in the context of finite-population inference from sample survey data and to the "ancillarity paradoxes" of Brown (1990) and Foster and George (1996) mentioned earlier. Indeed, the development of Brown (1990) suggests that any estimator that is unconditionally admissible for squared error loss will fail to be SFB and hence will violate the LP.

It is often stated that Bayesian inference always satisfies the likelihood principle. This is correct only when the prior does not depend on the experiment. Consider, for example, an observer with a prior π that makes θ and f dependent. Now suppose that this observer learns the true value f_0 of f. Then his posterior distribution of θ and ψ will depend on f_0 and thus violate the LP. Note that f_0 indexes the experiment being performed. Since, for any two experiments f_0 and f_0^*, the likelihood ratio $\mathcal{L}(\theta, f_0)/\mathcal{L}(\theta, f_0^*)$ is not a

function of θ, any estimator that depends on f_0 violates LP. This result does not contradict Birnbaum's theorem, because the observer's inferences also violate the CP because with this prior, knowledge of which experiment was actually performed (i.e., the true f_0 that generated the data) contains information concerning θ. To see this, consider the extreme case where the observer gets no data but learns f_0. Then observer's posterior for θ will be the conditional prior $\pi(\theta|f_0)$. However, if the process determining which experiment was chosen had been superseded and the experiment had instead been chosen by a coin flip, then the posterior for θ would be the marginal prior $\pi(\theta)$. In other words, the observer's inferences are (correctly) influenced by how the observer got to learn f_0.

The example we presented in this section is important for two reasons. First, as we noted earlier, a version of this problem actually arises in the problem of estimating treatment effects in randomized trials when the randomization probabilities depend on observed covariates. Second, the example illustrates the general point that good frequentist performance and the LP can be in severe conflict in the sense that any procedure with good frequentist properties must violate the LP. In the future, we believe that more attention should be directed to examining foundational principles in infinite-dimensional settings.

REFERENCES

Aldrich, J. (1997), "R. A. Fisher and the Making of Maximum Likelihood 1912–1922," *Statistical Science*, 12, 162–176.

Amari, S. (1982), "Geometrical Theory of Asymptotic Ancillarity and Conditional Inference," *Biometrika*, 69, 1–17.

Barndorff-Nielsen, O. (1978), *Information and Exponential Families in Statistical Theory*, New York: Wiley.

——— (1983), "On a Formula for the Distribution of the Maximum Likelihood Estimator," *Biometrika*, 70, 343–365.

Barron, A. (1988), "The Exponential Convergence of Posterior Probabilities With Implications for Bayes Estimators of Density Functions," unpublished manuscript.

Barron, A., Schervish, M., and Wasserman, L. (1999), "Consistency of Posterior Distributions in Nonparametric Problems," *The Annals of Statistics*, 27, 536–561.

Berger, J., and Wolpert, R. (1984), *The Likelihood Principle*, Hayward, CA: Institute for Mathematical Statistics.

Bickel, P., Klaassen, C., Ritov, Y., and Wellner, J. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore: Johns Hopkins University Press.

Birnbaum, A. (1962), "On the Foundations of Statistical Inference" (with discussion), *Journal of the American Statistical Association*, 57, 269–306.

Brown, L. D. (1967), "The Conditional Level of Student's t Test," *The Annals of Mathematical Statistics*, 38, 1068–1071.

——— (1978), "A Contribution to Kiefer's Theory of Conditional Confidence Procedures," *The Annals of Statistics*, 6, 59–71.

——— (1990), "An Ancillary Paradox which Appears in Multiple Linear Regression," *The Annals of Statistics*, 8, 471–493.

Buehler, R. J. (1959), "Some Validity Criteria for Statistical Inference," *The Annals of Mathematical Statistics*, 30, 845–863.

Buehler, R. J., and Feddersen, A. P. (1963), "Note on a Conditional Property of Student's t," *The Annals of Mathematical Statistics*, 34, 1098–1100.

Casella, G. (1987), "Conditionally Acceptable Recentered Set Estimators," *The Annals of Statistics*, 15, 1363–1371.

Cox, D. R. (1958), "Some Problems Connected With Statistical Inference," *The Annals of Mathematical Statistics*, 29, 357–372.

——— (1980), "Local Ancillarity," *Biometrika*, 67, 279–286.

——— (1988), "Some Aspects of Conditional and Asymptotic Inference:

A Review," *Sankhya*, 50, 314–337.

——— (1993), "An Analysis of Bayesian Inference for Nonparametric Regression," *The Annals of Statistics*, 21, 903–923.

Cox, D. R., and Hinkley, D. (1974), *Theoretical Statistics*, London: Chapman and Hall.

Cox, D. R., and Reid, N. (1987), "Parameter Orthogonality and Approximate Conditional Inference," *Journal of the Royal Statistical Society*, Ser. B, 49, 1–18.

Diaconis, P., and Freedman, D. (1986), "On the Consistency of Bayes Estimates," *The Annals of Statistics*, 14, 1–26.

——— (1990), "On the Uniform Consistency of Bayes Estimates for Multinomial Probabilities," *The Annals of Statistics*, 18, 1317–1327.

——— (1993), "Nonparametric Binary Regression: A Bayesian Approach," *The Annals of Statistics*, 21, 2108–2137.

——— (1997), "Consistency of Bayes Estimates for Nonparametric Regression: A Review," in *Festschrift for L. Le Cam*, 157–165.

DiCiccio, T. (1986), "Approximate Conditional Inference for Location Families," *Canadian Journal of Statistics*, 14, 5–18.

Doob, J. L. (1949), "Application of the Theory of Martingales," in *Le Calcul des Probabilités et ses Applications*, Colloques Internationaux du Centre National de la Recherche Scientifique, Paris, pp. 23–27.

de Finetti, B. (1974, 1975), *Theory of Probability, Vols. I and II*, trans. by A. F. M. Smith and A. Machi, New York: Wiley.

Edwards, A. W. F. (1972), *Likelihood*, Cambridge, U.K.: Cambridge University Press.

——— (1997), "What did Fisher Mean by "Inverse Probability?," *Statistical Science*, 12, 177–184.

Efron, B., and Hinkley, D. (1978), "Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information," *Biometrika*, 65, 457–482.

Evans, M., Fraser, D. A. S., and Monette, G. (1986), "On Principles and Arguments to Likelihood" (with discussion), *Canadian Journal of Statistics*, 14, 181–199.

Fienberg, S. (1997), "Introduction to R. A. Fisher on Inverse Probability and Likelihood," *Statistical Science*, 12, 161.

Fisher, R. A. (1921), "On the 'Probable Error' of a Coefficient of Correlation Deduced From a Small Sample," *Metron*, I. part 4, 3–32.

——— (1925), "Theory of Statistical Estimation," *Proceedings of the Cambridge Philosophical Society*, 22, 700–725.

——— (1934), "Two New Properties of Mathematical Likelihood," *Proceedings of the Royal Society*, Ser. A, 144, 285–307.

——— (1956), *Statistical Methods and Scientific Inference*, Edinburgh: Oliver and Boyd.

Foster, D. P., and George, E. I. (1996), "A Simple Ancillarity Paradox," *Scandinavian Journal of Statistics*.

Fraser, D. (1977), "Confidence, Posterior Probability and the Buchler Example," *The Annals of Statistics*, 5, 892–898.

Freedman, D. (1963), "On the Asymptotic Behavior of Bayes's Estimates in the Discrete Case," *The Annals of Mathematical Statistics*, 34, 1386–1403.

——— (1965), "On the Asymptotic Behavior of Bayes Estimates in the Discrete Case, II," *The Annals of Mathematical Statistics*, 36, 454–456.

Freedman, D., and Diaconis, P. (1983), "On Inconsistent Bayes Estimates in the Discrete Case," *The Annals of Statistics*, 11, 1109–1118.

Freedman, D. A., and Purves, R. A. (1969), "Bayes Methods for Bookies," *The Annals of Mathematical Statistics*, 40, 1177–1186.

Godambe, V. P., and Thompson, M. E. (1976), "Philosophy of Survey-Sampling Practice" (with discussion), in *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science, Vols. I, II, and III*, eds. W. L. Harper and A. Hooker, Dordrecht: D. Reidel.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999), "Posterior Consistency of Dirichlet Mixtures in Density Estimation," *The Annals of Statistics*, 27, 143–158.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. (1998), "Rates of Convergence of Posteriors," technical report, Free University, Amsterdam.

Heath, D., and Sudderth, W. (1978), "On Finitely Additive Priors, Coherence, and Extended Admissibility," *The Annals of Statistics*, 6, 333–345.

——— (1989), "Coherent Inference From Improper Priors and From Finitely Additive Priors," *The Annals of Statistics*, 17, 907–919.

Jeffreys, H. (1961), *The Theory of Probability*, Oxford, U.K.: Clarendon Press.

Kiefer, J. (1977), "Conditional Confidence Statements and Confidence Estimators" (with discussion), *Journal of the American Statistical Associ-*

ation, 72, 789–827.

Lehmann, E. L. (1986), *Testing Statistical Hypotheses* (2nd ed.), New York: Wiley.

Lindley, D. V., and Phillips, L. D. (1976), "Inference for a Bernoulli Process (A Bayesian view)," *The American Statistician*, 30, 112–119.

Lindsay, B. (1980), "Nuisance Parameters, Mixture Models, and the Efficiency of Partial Likelihood Estimators," *Philosophical Transactions of the Royal Society*, 296, 639–665.

Ramsey, F. (1931), *The Foundations of Mathematics and Other Logical Essays*, Paterson, NJ: Littlefield-Adams.

Regazzini, E. (1987), "De Finetti's Coherence and Statistical Inference," *The Annals of Statistics*, 15, 845–864.

Robins, J. M., and Morgenstern, H. (1987), "The Foundations of Confounding in Epidemiology," *Computers and Mathematics with Applications*, 14, 869–916.

Robins, J. M., and Ritov, Y. (1997), "A Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semiparametric Models," *Statistics in Medicine*, 16, 285–319.

Robins, J. M., Rotnitzky, A., and van der Laan, M. (2000), Comment on "On Profile Likelihood," by S. A. Murphy and A. W. van der Vaart, *Journal of the American Statistical Association*, 95, 477–482.

Robinson, G. K. (1976), "Properties of Student's *t* and of the Behrens–Fisher Solution to the Two Means Problem," *The Annals of Statistics*, 5, 963–971.

———— (1979), "Conditional Properties of Statistical Procedures," *The Annals of Statistics*, 7, 742–755.

Savage, L. J. (1954), *The Foundations of Statistics*, New York: Wiley.

Schervish, M. (1995), *Theory of Statistics*, New York: Springer-Verlag.

Schwartz, L. (1960), "Consistency of Bayes Procedures," Ph.D. dissertation, University of California.

———— (1965), "On Bayes Procedures," *Z. Wahrsch. Verw. Gebiete*, 4, 10–26.

Severini, T. (1993), "Local Ancillarity in the Presence of a Nuisance Parameter," *Biometrika*, 80, 305–320.

Shen, X., and Wasserman, L. (1998), "Rates of Convergence of Posterior Distributions," Technical Report 678, Carnegie Mellon University, Statistics Dept.

Sprott, D. A. (1975), "Marginal and Conditional Sufficiency," *Biometrika*, 62, 599–606.

Sweeting, T. (1992), "Asymptotic Ancillarity and Conditional Inference for Stochastic Processes," *The Annals of Statistics*, 20, 580–589.

Welch, B. L., and Peers, H. W. (1963), "On Formulae for Confidence Points Based on Integrals of Weighted Likelihoods," *Journal of the Royal Statistical Society*, Ser. B, 25, 318–329.

Zhao, L. (1993), "Frequentist and Bayesian Aspects of Some Nonparametric Estimation Problems," Ph.D. thesis, Cornell University.

———— (1998), "A Hierarchical Bayesian Approach in Nonparametric Function Estimation," technical report, University of Pennsylvania, Wharton School.

The End of Time Series

V. SOLO

1. INTRODUCTION

Just as physicists have spent the last three-quarters of a century working out the consequences of the quantum mechanics paradigm established by the late 1920s, so have statisticians spent the last half century working out the consequences of the statistical paradigm assembled by the end of World War II.

As a branch of statistics, time series has always occupied a special place. Not only is it mostly practiced by nonstatisticians, but, as a consequence, most of the innovation has come from outside statistics. But statisticians have nevertheless played a disproportionate role given their relatively small numbers. A list of areas making a significant use of time series methods includes communications (Haykin 1983), Control Engineering (Ljung 1987), econometrics (Hamilton 1994), geophysics (Aki and Richards 1980), meteorology (Daley 1991), optical signal processing (Papoulis 1968), signal processing (Kay 1988), and radio astronomy (Steward 1987). In fact, any applied discipline that makes heavy use of Fourier methods is likely to have some need for time series.

Before plunging into the details, I outline the dimension of my discussion space. Some elementary partitions are obvious: discrete and continuous time, stationary and nonstationary, scalar and vector, linear and nonlinear, time domain and frequency domain, and parametric and nonparametric.

Finally, I observe, less obviously, that the problems of time series analysis resolve mostly into modeling and/or filtering. Because each of the resulting 128 cells will have several subheadings, it is clear that the discussion must be selective! Some of these cells have been well worked out, others have been barely touched, and others, though much studied, have thrown up problems that have long resisted solution. Some other topics, such as wavelets (being time-frequency methods) and semiparametrics, require partitioning to allow interaction; this results in 2,187 cells!

2. A BRIEF HISTORY OF TIME (SERIES)

The period of interest may be roughly divided into two subperiods: 1945–1975 and 1975–1999. In the first subperiod, the following developments can be identified:

1. Development of basic filtering and smoothing algorithms, including fast versions. Thus we have Levinson's scalar algorithm (Bartlett 1955; Durbin 1960, who found a variance update missed by Levinson; Wiener 1949, appendix); the vector version of Levinson's algorithm (Robinson 1967; Whittle 1963a); the Kalman filter (Kalman 1960) based on state-space representations; Spectral factorization for prediction filter design is closely related—it is the infinite-dimensional version of the finite-data Cholesky factorization (of the covariance matrix inverse) achieved by the Kalman filter. It is quite remarkable that even today, some

V. Solo is Professor of Electrical Engineering, University of New South Wales, Sydney, NSW 2052, Australia (E-mail: *vsolo@syscon.ee.unsw.edu.au*). This work was completed while the author was Professor of Statistics at Macquarie University, Sydney, Australia.