The Foundations of Statistics: A Vignette

James Robins[1] and Larry Wasserman[2]
Harvard School of Public Health and Carnegie Mellon University
August 22, 1999

This vignette reviews some issues in the Foundations of Statistics. Our review is very selective, with particular emphasis on the role of the likelihood and conditioning. Specifically, we review the likelihood principle, conditional inference and coherence arguments. We close with a brief look at some modern theory which examines the role of the likelihood principle in nonparametric problems.

1. Introduction.

Statistics is intertwined with science and mathematics but it is a subset of neither. The "Foundations of Statistics" is the set of concepts that makes Statistics a separate and distinct field. For example, arguments for and against conditioning on ancillaries are purely statistical in nature; mathematics and probability do not inform us on 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.

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 two, 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 two we discuss the likelihood function, the likelihood principle, the conditionality principle and the sufficiency principle. In Section three

[1]Research supported by N.I.H.
[2]Research supported by N.I.H. grant R01-CA54852-01 and NSF grants DMS-9303557 and DMS-9357646. The authors thank Isabella Verdinelli for many helpful suggestions.

we briefly review conditional inference. In Section four we discuss Bayesian inference and coherence. Section five contains a brief look at some newer foundational work which 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 we have n i.i.d. 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) 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 Fienberg (1997), Aldrich (1997) and Edwards (1997) for a discussion 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 $\hat{\theta}_n$ – 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 $C_n = \{\theta; \mathcal{L}_n(\theta)/\mathcal{L}_n(\hat{\theta}) \leq c\}$ gives the asymptotically shortest, $1-\alpha$ confidence set if c is chosen appropriately.

Clearly, the likelihood function is a crucial part of many statistical analyses. More controversial is the role of the "likelihood principle." The likelihood principle (LP) says that two experiments that yield proportional likelihoods 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 defined above. Suppose two different experiments yield proportional likelihoods. Then the form of the interval C_n would be the same in the two experiments but the coverage probability ascribed to C_n could be different, thus violating the LP.

Birnbaum (1962) caused quite a stir when he 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 is contained in Berger and Wolpert (1984). This and the next section draw heavily from that monograph.

The conditionality principle (Cox, 1958) says that if we decide which of two experiments to do by the flip of a 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 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 we 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, our inferences should be identical to the inferences we would make if E_1 was performed without first flipping the coin (and similarly for E_2). (In the next section, we shall see that 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 each to be 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) have since shown that LP follows from a slightly stronger version of CP alone.

Most frequentist methods violate LP. For example, standard Neyman-Pearson tests and confidence intervals blatantly violate LP. The only methods that obey LP are strict Bayesian methods and methods that use only the observed likelihood function (Edwards 1972) without using the sampling properties of the likelihood function or associated likelihood-based point and interval estimators.

The monograph by Berger and Wolpert (1984) contains a lively discussion including supporters and detractors. Let us add our own point of view. The CP seems compelling in the above simple example whenever knowledge of which experiment was performed provides no additional information about the parameter beyond that contained in the data. In Section five we show that the CP is less compelling in high dimensional models.

3. Conditioning, Ancillarity and Relevant Subsets

In the last 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 that are 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 i.i.d 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 "confidence set" though, unlike the typical confidence set, it contains only a single point. This set is in fact a 75 per cent confidence set. In other words, $P_\theta(\theta \in C) = 0.75$ for all θ. But when $Y_1 \neq Y_2$ we are certain that $\theta \in C$ and when $Y_1 = Y_2$, it seems intuitively clear that we should have 50 per cent confidence that $\theta \in C$. 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 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 are discussed in Fisher (1956).

An ancillary statistic is a statistic whose distribution does not depend on θ. In the example above, $S = |Y_1 - Y_2|$ is ancillary and it seems quite reasonable to report a confidence of 1 when $S = 1$ and a confidence of 0.50 when $S = 0$. Specifically, we report $P_\theta(\theta \in C|S = 1) = 1$ and $P_\theta(\theta \in C|S = 0) = 0.5$. The coin flip in the mixed experiment described in Section two is another example of an ancillary statistic.

Consider a $1-\alpha$ confidence set $C(Y)$, i.e. $P_\theta(\theta \in C(Y)) = 1-\alpha$ for all θ. We say that B is a relevant (or recognizable) subset if there is some $\epsilon > 0$ such that either, $P_\theta(\theta \in C(Y)|Y \in B) \leq (1-\alpha) - \epsilon$ for all θ or $P_\theta(\theta \in C(Y)|Y \in B) \geq (1-\alpha) + \epsilon$ 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 since 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 include Brown (1967, 1978), Casella (1987), Cox (1958, 1980), Buehler (1959, 1976), 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 upon which to condition. One way to extend the applicability of conditional inference is to use approximate conditional inference in which one conditions on a statistic which is asymptotically ancillary; see Amari (1982), Barndorff-Nielsen (1983), Cox (1988), Cox and Reid (1987), DiCiccio (1986), Efron and Hinkley (1978), Robins and Morgenstern (1987), Sweeting (1992), Severini (1993), and references therein.

Even when there is an obvious statistic to condition on, it is still not clear that we should indeed condition. Brown (1990) gives an example in which an estimator is admissible conditional on an ancillary statistic but is unconditionally inadmissible; see also Foster and George (1996). The conclusion we can draw from this is that, while conditioning on an ancillary seems appealing in some examples, it is not obvious that we should always condition on an ancillary. We discuss this further in Section 5.

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, since they are meant to capture what a coherent (i.e. sensible) inference is. Usually, these arguments lead to conclusions of the form: 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 since 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 its 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, for example, De Finetti (1974, 1975), Freedman and Purves (1969), Heath and Sudderth (1978, 1989), Jeffreys (1961), Ramsey (1930), Regazzini (1987) and 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 upon observing $Y = y$ which have a strictly positive expected payoff. Heath and Sudderth prove 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 inference 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, there are many statisticians who would choose the latter. The issue is not mathematical in nature. The question is whether one finds coherence or correct coverage more important.

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

(1) $\overline{\theta}_n - \hat{\theta} = O_P(n^{-1})$;

(2) If C_n is an appropriate $1-\alpha$ posterior region i.e. $\int_{C_n} Q(d\theta|Y^n) = 1-\alpha$, then C_n has frequentist coverage $1-\alpha+O(n^{-1})$.

(3) If N is any 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 at rate $n^{-1/2}$. In other words, if $A_n \to \infty$ and θ_0 denotes the true value, then $Q(\{\theta\,;\,|\theta-\theta_0| \geq A_n n^{-1/2}\}|Y^n) = o_P(1)$.

In words, (1) the maximum likelihood estimate and posterior mean are close, (2) Bayesian posterior intervals and confidence intervals agree approximately, (3) the posterior is consistent and (4) the posterior converges at the same rate as the maximum likelihood estimate.

In nonparametric models, the situation is less clear. Consistency is sometimes attainable and sometimes not: see Diaconis and Freedman (1986, 1990, 1993, 1997), Freedman (1963, 1965), Freedman and Diaconis (1983), Ghosal, Ghosh and Ramamoorthi (1999), Barron (1988), Barron, Schervish and Wasserman (1999), Doob (1949), and 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) and Zhao (1993, 1998). The issue of matching posterior probability and frequentist coverage has received less attention. While there are some negative results in Freedman (1999) and Cox (1993), this topic remains mostly unexplored territory.

5. A Look to the Future: Foundations in Nonparametric Inference.

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 re-think Foundations in light of these methods? Here we argue that the answer is yes. We summarize an example which is discussed in great detail by Robins and Ritov (1997). To keep things brief and simple, we give a telegraphic version and omit the details. The reader should see the Robins-Ritov paper for a full discussion. Although it will not be immediately obvious, the example stems from a real problem, namely, the analysis of treatment effects in randomized trials.

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 $\sup_{x \in (0,1)^k} |\theta(x)| \leq M$ for some 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 $\mathbf{X} = \{X_i; i = 1, ..., n\}$ denote the observed $X_i's$. When f_0 is known, $\mathbf{X}$ is ancillary. When f_0 is unknown, $\mathbf{X}$ is 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).

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 ancillarity. When f_0 is unknown, Robins and Ritov (1997) have shown that no uniformly consistent estimator of ψ_0 exists. 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 $\overline{V} = n^{-1}\sum_{i=1}^n V_i = n^{-1}\sum_{i=1}^n Y_i/f_0(X_i)$ 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, that is, when k is large.)

This result has implications for many common inferential methods. In particular, standard likelihood and Bayesian methods will fail in this model. 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 be the known $f_0 \in \mathcal{F}_X$ that generated the data. We call such methods *strict factorization based* (SFB) methods. Indeed, any inference method that satisfies the likelihood principle is SFB. Robins and Ritov (1997) show 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 are not only asymptotic. Any interval estimator that is not a function of f_0 will not be "valid". By valid we mean that under all $(\theta_0, f_0) \in \Theta \times \mathcal{F}_X$ (i) the coverage is at least $(1-\alpha)$ at each sample size n and (ii) the expected length goes to zero 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 $\overline{V} \pm dn^{-1/2}$ 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 criticisms made of likelihood based inference by Godambe (1970) in the context of finite population inference from sample survey data.

It is often stated that Bayesian inference always satisfies the likelihood principle. However this is incorrect. 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. However, this does not contradict Birnbaum's theorem since 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 natural processes determining which experiment was performed were superseded and the experiment to be performed was instead chosen by a coin flip, the posterior for θ would be the marginal prior $\pi(\theta)$. In other words, the inferences are (correctly) influenced by how it is the observer got to learn f_0.

The example we have 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 infinite dimensional problems 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 nonparametric 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*. Wiley: New York.

Barndorff-Nielsen, O. (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. To appear: *Ann. Statist.*

Berger, J. and Wolpert, R. (1984). *The Likelihood Principle*. IMS lecture notes. Hayward, CA.

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.

Brown, L.D. (1978). A contribution to Kiefer's theory of conditional confidence procedures. *The Annals of Statistics*, 6, 59-71.

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.

Cox, D.R. (1980). Local ancillarity. *Biometrika*, 67, 279-286.

Cox, D. R. (1988). Some aspects of conditional and asymptotic inference: A review. *Sankhya*, 50, 314-337.

Cox, D. (1993). An analysis of Bayesian inference for nonparametric regression. *The Annals of Statistics*, 21, 903-923.

Cox, D. R and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistica Society*, 49, 1-18.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *The Annals of Statistics*, 14, 1-26.

Diaconis, P. and Freedman, D. (1990). On the uniform consistency of Bayes estimates for multinomial probabilities. *The Annals of Statistics*, 18, 1317-1327.

Diaconis, P. and Freedman, D. A. (1993). Nonparametric binary regression: A Bayesian approach. *The Annals of Statistics*, 21, 2108-2137.

Diaconis, P. and Freedman, D. A. (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, 23-27.

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

Edwards, A.W.F. (1972). *Likelihood.* C.U.P. Cambridge.

Edwards, A.W.F. (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.

Fisher, R.A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22, 700-725.

Fisher, R.A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society, A*, 144, 285-307.

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

Foster, D.P. and George, E.I. (1996). A simple ancillarity paradox. *Scandinavian Journal of Statistics.*

Freedman, D. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case *The Annals of Mathematical Statistics*, 34, 1386-1403.

Freedman, D. A. (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. (1970). Foundations of survey-sampling. *The American Statistician*, 24, 33-38.

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.

Heath, D. and Sudderth, W. (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.* Clarendon Press: Oxford.

Kiefer, J. (1977). Conditional confidence statements and confidence estimators (with discussion). *Journal of the American Statistical Association*, 72, 789-827.

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

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. In *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.

Robinson, G.K. (1976). Properties of Student's t and of the Behrens-Fisher solution to the two means problem. *Annals of Statistics*, 5, 963-971.

Robinson, G.K. (1979). Conditional properties of statistical procedures. *Annals of Statistics*, 7, 742-755.

Rubin, D. (1974). Estimating the causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688-701.

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

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

Schwartz, L. (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, Statistics Department, Carnegie Mellon University.

Sweeting, T. (1992). Asymptotic ancillarity and conditional inference for stochastic processes.

The Annals of Statistics, 20, 580-589.

Zhao, L. (1993). Frequentist and Bayesian aspects of some nonparametric estimation problems. Ph.D. thesis. Cornell University.

Zhao, L. (1998). A hierarchical Bayesian approach in nonparametric function estimation. Technical report, Wharton School, University of Pennsylvania.