Statistical Science
2010, Vol. 25, No. 2, 158–161
DOI: 10.1214/10-STS308A
Main article DOI: 10.1214/09-STS308

Comment: The Need for Syncretism in Applied Statistics

Sander Greenland

It is an honor to comment on Prof. Efron's latest contribution to the merging of frequentist and Bayesian thinking into a harmonious (even if not strictly coherent) statistical viewpoint. I will review my thinking along those lines and some inspirations for it. I agree with most of Dr. Efron's views expressed here and in Efron (2005), with these important exceptions: First, I disagree that frequentism has supplied a good set of working rules. Instead, I argue that frequentism has been a prime source of reckless overconfidence in many fields (especially but not only in the form of 0.05-level testing; see Rothman, Greenland and Lash, 2008, Chapter 10 for examples and further citations). I also disagree that Bayesians are more aggressive than frequentists in modeling. The most aggressive modeling is that which fixes unknown parameters at some known constant like zero (whence they disappear from the model and are forgotten), thus generating overconfident inferences and an illusion of simplicity; such practice is a hallmark of conventional frequentist applications in observational studies.

As working rules, the problem with conventional methods lies not so much with frequentism, but rather with frequentist tools for designed experiments being misapplied to observational data (Greenland, 2005a). Bayesians can and do misapply their methods similarly; they just haven't been given as much opportunity to do so. Conversely, many frequentist as well as Bayesian tools for observational studies have been developed, especially for sensitivity analysis. But the overconfidence problem has been perpetuated by the ongoing concealment of unbelievable point-mass priors within models in order to maintain frequentist identification of target parameters.

The problem can be addressed by sacrificing identification and replacing bad modeling assumptions with explicit and reasonable priors (Gustafson, 2005; Greenland, 2005a, 2009a). Perhaps ironically, frequentist thought experiments and simulations can then provide both contextual and frequentist diagnostics (Rubin, 1984; Greenland, 2006; Gustafson and Greenland, 2009). Thus, frequentist thinking can address Bayesian overconfidence just as Bayesian thinking can address frequentist overconfidence. Hence I would strengthen Box's plea for ecumenism (Box, 1983) into an imperative to fuse Bayesian and frequentist concepts and methods in statistical inference—and in teaching as well. This theme is far from new (e.g., besides Box, see Good, 1983; Diaconis and Freedman and discussants, 1986; Samaniego and Reneau, 1994), yet it has barely touched everyday teaching and practice. In this case (unlike many) that is not because of software limitations; in fact, for the bulk of applications the same software can be used for both frequentist and Bayesian calculations (Greenland, 2007, 2009a).

HIERARCHICAL MODELING: WHERE PRIORS AND FREQUENCIES MEET

Bayesian and frequentist ideas intertwine in hierarchical modeling (Efron's Section 9), which encompasses both Bayesian and empirical-Bayes approaches (Good, 1983, 1987) as well as other shrinkage techniques. Efron and Morris (1973, 1975) were among the earliest to demonstrate convincingly that hierarchical models offered practical as well as theoretical advantages for data analysis. Their writings (along with those of Jack Good, George Box and Edward Leamer) inspired my applications of hierarchical modeling and Bayesian methods in epidemiology, where the hierarchy levels are naturally determined by physical structures and observation processes.

As an example, in nearly all observational studies of nutrient effects, individual risks are regressed directly on nutrient intakes calculated from food intakes. This conventional model makes no further use of the food intakes, and so assumes implicitly that foods have no effect on risk beyond their calculated nutrient content. This is an unsupported and very doubtful assumption. A more realistic model allows food effects beyond

S. Greenland is Professor, Department of Epidemiology and Department of Statistics, University of California, Los Angeles, California 90095-1772, USA (e-mail: lesdomes@ucla.edu)

measured-nutrient content. However, the resulting two-level hierarchical model is not identified without a prior because nutrient intakes are linear functions of food intakes (making nutrient and food intakes completely collinear). Using any contextually defensible prior reveals that the conventional analysis generates overconfident inferences, both in the Bayesian sense of overstating information (Greenland, 2000), and also in the frequentist sense of producing interval undercoverage (Gustafson and Greenland, 2006). That overconfidence may explain the rather embarrassing track record of nutritional epidemiology when compared against clinical trials (Lawlor et al., 2004). Ecologic analyses provide other examples in which use of the natural hierarchical structure with explicit priors is needed to avoid overconfidence (Wakefield, 2009).

In this work, I have come to appreciate that a simultaneously Bayesian and frequentist viewpoint is essential for a credible analysis of observational data. I must be at least informally Bayesian, knowing that there is no contextual credibility without consideration and use of prior information, especially in model specification. But I should also be at least informally frequentist, knowing that priors should be weighted lightly unless they derive from statistical observations such as frequencies in partially exchangeable past experience (e.g., surveys) or classical measurement processes (e.g., laboratory determinations). Most of all, I should not rigidly adhere to ideologies or models, especially when a clash between my prior and my likelihood function shows that my understanding of the situation is more deficient than I initially thought (Box, 1980, 1990).

PRIORS: EVERYBODY USES THEM (BUT MOST CALL THEM "MODELS")

As Efron illustrates in Section 4, all analyses labeled as frequentist are built on priors, although these priors are called "models," which avoids the controversies associated with overtly Bayesian analysis (Leamer, 1978; Box, 1980, 1983). Even the simplest randomized-trial analysis is based on a model, namely the prior belief that treatment was randomized fairly, and the reported subjects actually exist. As numerous cases of fraud demonstrate, that belief may be mistaken more often than those receiving medical treatment would like to think (e.g., see Greenland, 2009b).

Labeling assumptions and models as prior beliefs might better alert us to the act of faith involved in their use. As Box (1980) said

> I believe that it is impossible logically to distinguish between model assumptions and the prior distribution of the parameters. The model is the prior in the wide sense that it is a probability statement of all the assumptions currently to be tentatively entertained a priori. On this view, traditional sampling theory was of course not free from assumptions of prior knowledge. Instead it was as if only two states of mind had been allowed: complete certainty or complete uncertainty.

I have grown increasingly uncomfortable with the convention of failing to label models as priors. It encourages the use of arbitrary constraints, and questions constraints only if the analysis data (the direct evidence) can reveal departures—even though studies are not designed with anywhere near sufficient power to reveal all important model violations. The representation of modeling constraints in belief networks (Madigan, Mosurski and Almond, 1997) can aid in the display of these constraints as imposed beliefs and thus expose implausible aspects of the model, although of course it cannot address data limitations. Yet single datasets are often too limited to tell us much about either the effects under study or our models (Robins and Greenland, 1986)—at least if we do not impose a hoard of dubious independence constraints that amount to point-mass priors with no supporting data.

Additivity in generalized linear models is an example: with n covariates, additivity sets all orders of product terms ("interactions") among them to zero, and is equivalent to using a point mass at zero for the joint prior on these terms. Entering the few "significant" two-way products hardly makes a dent in this set of constraints if $n > 5$; yet $n > 8$ is common and $n > 20$ not unusual. Arbitrary additive constraints can be relatively harmless when estimating a population-average effect, because the specification error they entail may average out in much the way random residual error does (Greenland and Maldonado, 1994). But the constraints can be deadly when used for individual (clinical) risk prediction, as adverse drug interactions demonstrate.

Hierarchical methods offer one way to relax additivity constraints in a controlled fashion, by including all or many products but shrinking their estimates toward zero or a second-level structure (Wakefield, De Vocht and Hung, 2010). More generally, we can expand an unrealistic conventional model by embedding it in a richer, more realistic hierarchical model, then shrink

estimates from the latter using prior distributions. Aspects of these distributions may be chosen to improve frequency performance in high-dimensional problems, but such methods do not preclude the use of prior information to judge those and other aspects of the formal prior distribution.

THE NEED FOR EXPLICIT PRIORS IN OBSERVATIONAL STUDIES

My discomfort with conventional treatments of modeling has increased knowing that observational data analysis can identify causal effects *only* by using indirect evidence, no matter how large the dataset or how informed by past observational data. This is the usual situation in epidemiology, where confounders, selection-probability ratios, or valid exposure measurements are unavailable for analysis (Greenland, 2005a; Gustafson, 2005; Rothman, Greenland and Lash, 2008, Chapter 19; Lash, Fox and Fink, 2009). The problem is a variant of the nonidentifiability of a regression coefficient when some regressors are latent (Leamer, 1974). In these cases a credible formal analysis must introduce proper priors in place of overconfident identifying constraints.

Use of identified regression models as sources of effect estimates corresponds to a multidimensional point prior that says there is no uncontrolled confounding or selection bias, and that measurements (including validation measurements) were accurate or at least reliable for life histories. Taken jointly, these assumptions are absurd in topics like nutritional and "lifestyle" epidemiology. But relaxing these silly and harmful assumptions leads to a realm where most Bayesians as well as frequentists fear to tread: Specification of prior distributions that cannot be effectively checked or updated with the analysis data.

When the scientific validity of each analysis hinges on extensive and untestable prior specifications, an analysis can be no more than a rough guess about a vast unknown, and represents but one element in a sensitivity analysis (Greenland, 2005b). This is true even of a formal sensitivity analysis, which is limited to examining a few parameters lest it become unintelligible. In this reality, the importance of specific models and priors should be de-emphasized in favor of providing a framework for sensitivity analysis across plausible models and priors. Accuracy of computation becomes secondary to prior specification, which is too often neglected under the rubric of "objective Bayes" (a.k.a. "please don't bother me with the science" Bayes).

There is simply no point in trying to do well at all conceivable parameter values given the model when the model embedding the parameter has already imposed doubtful point constraints. Hence I have sought approaches in which informative priors are central. Good (1983) provided the key ingredients: Priors can be transformed into penalty functions, which can then be transformed into "prior data" that generate the penalties as log-likelihood contributions. This transformation allows evaluation of prior-knowledge claims in a currency familiar to the subject-matter expert, as well as use of familiar and rapid fitting methods for basic models (Bedrick, Christensen and Johnson, 1996, 1997; Greenland, 2006, 2007, 2009a, 2009c).

Note that conversion of priors to prior data does *not* require conjugacy; it only requires that the penalties have representations as transformed likelihoods from a series of observations or experiments. The credibility of the prior may be questioned if such a representation is absent, arcane, or absurd. Evaluation of priors in terms of equivalent data is particularly illuminating in human-subject fields, where data are expensive and hence sparse. Here, strong priors may be seen as claiming access to a volume of data that does not exist, thus casting doubt on prior assertions of some experts (Higgins and Spiegelhalter, 2002; Greenland, 2006).

When priors (the indirect evidence) are recalibrated to match the frequentist outputs of reasonably sized thought experiments, the combined evidence will often be too limited to distinguish among the effect sizes at issue (Greenland, 2009c). This is unwelcome news to some colleagues, albeit no news to others. Regardless, the future of indirect evidence should be recognition for what it is: Omnipresent and essential for any inference beyond "more research is needed" (which may the strongest conclusion we can hope to wrest from most studies, albeit not always justifiable in economic terms).

Thus I would conclude by echoing Efron: Whether Bayesian, frequentist, ecumenic, or syncretic, statisticians need to become better at creating and evaluating contextually informed models—which include both well-informed prior distributions and sensible qualitative structures. It follows that statistical training should introduce informative-Bayesian methods in tandem with classical (and often destructive) frequentist methods, rather than as an afterthought or specialty topic. Data priors provide one easy and natural way to do so, displaying as they do the symmetry between indirect and direct evidence, and exposing priors to a new angle of criticism.

REFERENCES

BEDRICK, E. J., CHRISTENSEN, R. and JOHNSON, W. (1996). A new perspective on generalized linear models. *J. Amer. Statist. Assoc.* **91** 1450–1460. MR1439085

BEDRICK, E. J., CHRISTENSEN, R. and JOHNSON, W. (1997). Bayesian binomial regression: Predicting survival at a trauma center. *Amer. Statist.* **51** 211–218.

BOX, G. (1980). Sampling and Bayes inference in scientific modeling and robustness. *J. Roy. Statist. Soc. Ser. A* **143** 383–430. MR0603745

BOX, G. (1983). An apology for ecumenism in statistics. In: *Scientific Inference, Data Analysis, and Robustness* (G. Box, T. Leonard and C.-F. Wu, eds.) 51–84. Academic Press, New York. MR0772763

BOX, G. (1990). Comment on "The unity and diversity of probability" by Glen Shafer. *Statist. Sci.* **5** 448–449. MR1092984

DIACONIS, P. and FREEDMAN, D. (1986). On the consistency of Bayes estimates (with discussion). *Ann. Statist.* **14** 1–67. MR0829555

EFRON, B. (2005). Bayesians, frequentists, and scientists. *J. Amer. Statist. Assoc.* **100** 1–5. MR2166064

EFRON, B. and MORRIS, C. N. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130. MR0388597

EFRON, B. and MORRIS, C. N. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70** 311–319. MR0391403

GOOD, I. J. (1983). *Good Thinking*. Univ. Minnesota Press, Minneapolis. MR0723501

GOOD, I. J. (1987). Hierarchical Bayesian and empirical Bayesian methods (letter). *Amer. Statist.* **41** 92.

GREENLAND, S. (2000). When should epidemiologic regressions use random coefficients? *Biometrics* **56** 915–921.

GREENLAND, S. (2005a). Multiple-bias modeling for analysis of observational data (with discussion). *J. Roy. Statist. Soc. Ser. A* **168** 267–308. MR2119402

GREENLAND, S. (2005b). Contribution to discussion of Prentice, Pettinger, and Anderson. *Biometrics* **61** 920–921. MR2216182

GREENLAND, S. (2006). Bayesian perspectives for epidemiologic research. I. Foundations and basic methods (with comment and reply). *International Journal of Epidemiology* **35** 765–778.

GREENLAND, S. (2007). Bayesian perspectives for epidemiologic research. II. Regression analysis. *International Journal of Epidemiology* **36** 195–202.

GREENLAND, S. (2009a). Relaxation penalties and priors for plausible modeling of nonidentified bias sources. *Statist. Sci.* **24** 195–210. MR2655849

GREENLAND, S. (2009b). Dealing with uncertainty about investigator bias: Disclosure is informative. *Journal of Epidemiology and Community Health* **63** 593–598.

GREENLAND, S. (2009c). Bayesian perspectives for epidemiologic research. III. Bias analysis via missing-data methods. *International Journal of Epidemiology* **38** 1662–1673.

GREENLAND, S. and MALDONADO, G. (1994). The interpretation of multiplicative model parameters as standardized parameters. *Stat. Med.* **13** 989–999.

GUSTAFSON, P. (2005). On model expansion, model contraction, identifiability, and prior information: Two illustrative scenarios involving mismeasured variables (with discussion). *Statist. Sci.* **20** 111–140. MR2183445

GUSTAFSON, P. and GREENLAND, S. (2006). The performance of random coefficient regression in accounting for residual confounding. *Biometrics* **62** 760–768. MR2247204

GUSTAFSON, P. and GREENLAND, S. (2009). Interval estimation for messy observational data. *Statist. Sci.* **24** 328–342.

HIGGINS, J. P. T. and SPIEGELHALTER, D. (2002). Being skeptical about meta-analyses: A Bayesian perspective on magnesium trials in myocardial infarction. *International Journal of Epidemiology* **31** 96–104, appendix.

LASH, T. L., FOX, M. P. and FINK, A. K. (2009). *Applying Quantitative Bias Analysis to Epidemiologic Data*. Springer, New York.

LAWLOR, D. A., DAVEY SMITH, G., BRUCKDORFER, K. R., KUNDU, D. and EBRAHIM, S. (2004). Those confounded vitamins: What can we learn from the differences between observational versus randomized trial evidence? *Lancet* **363** 1724–1727.

LEAMER, E. E. (1974). False models and post-data model construction. *J. Amer. Statist. Assoc.* **69** 122–131.

LEAMER, E. E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. Wiley, New York. MR0471118

MADIGAN, D., MOSURSKI, K. and ALMOND, R. G. (1997). Graphical explanation in belief networks. *J. Comput. Graph. Statist.* **6** 160–181.

ROBINS, J. M. and GREENLAND, S. (1986). The role of model selection in causal inference from nonexperimental data. *American Journal of Epidemiology* **123** 392–402.

ROTHMAN, K. J., GREENLAND, S. and LASH, T. L., EDS. (2008). *Modern Epidemiology*, 3rd ed. Lippincott-Williams-Wilkins, Philadelphia.

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR0760681

SAMANIEGO, F. J. and RENEAU, D. (1994). Toward a reconciliation of the Bayesian and frequentist approaches to point estimation. *J. Amer. Statist. Assoc.* **89** 947–957. MR1294739

WAKEFIELD, J. (2009). Multi-level modelling, the ecologic fallacy, and hybrid study designs. *International Journal of Epidemiology* **38** 330–336.

WAKEFIELD, J., DE VOCHT, F. and HUNG, R. J. (2010). Bayesian mixture modeling of gene-environment and gene-gene interactions. *Genetic Epidemiology* **34** 16–25.