

**DISCUSSION OF “ANCILLARIES AND CONDITIONAL INFERENCE”
BY D.A.S. FRASER**

Rudolf Beran*
University of California, Davis

March 2003

The concepts of sufficiency, ancillarity, and conditional inference are parts of a classical statistical theory that treats data as a random sample from a probability model with relatively few parameters. In discussing Don Fraser’s paper, I will consider the place of these and related concepts in the evolution of statistics.

1. THE EVOLUTION OF STATISTICAL THEORY

Reliance on probability theory in statistical writing spans the spectrum from none, to fixed effects models, to random effects models, to Bayesian reasoning. One factor is the extent to which an author regards probability as a feature of the natural world. For a Bayesian, probability measures the strength of opinions, which are modeled by a sigma algebra. At the other end of the spectrum, illustrated by J. Tukey’s (1977) *Exploratory Data Analysis*, data-analytic algorithms are basic reality and probability models are hypothetical constructs.

A second factor is the technological environment in which an author is writing. Until the late 1950’s, the tools available to a statistician consisted of mathematics, logic, mechanical calculators, and simple computers. Because calculation was laborious, writers on statistical theory thought in terms of virtual data governed by probability models involving relatively few parameters. Indeed, the great intellectual advances made in probability theory during the twentieth century made this approach the technology of choice. Thus, the hotly debated statistical theories formulated in A. Wald’s (1950) *Statistical Decision Functions*, R. A. Fisher’s (1956) *Statistical Methods and Scientific Inference*, and L. J. Savage’s (1954) *The Foundations of Statistics* shared a common reliance on relatively simple probability models.

After 1960, results on weak convergence of probability measures provided the technology for major development of asymptotic theory in statistics. Notable achievements by 1970 included: (a) the clarification of what is meant by asymptotic optimality; (b) the understanding, through Le Cam’s work, that risks in simple parametric models can approximate risks in certain more general models; (c) the discovery of superefficient estimators whose asymptotic risk undercuts the information bound on sets of Lebesgue measure zero; and (d) the remarkable discovery, through the James-Stein estimator, that superefficient estimators for

* This research was supported in part by National Science Foundation Grant DMS 0300806.

parameters of sufficiently high dimension can dominate classical estimators globally. These findings set the stage for the vigorous subsequent development of robustness, of nonparametrics, and of biased estimators in models with many or an infinite number of parameters. Theoretical study of Efron's (1979) bootstrap benefited from the evolution in asymptotic theory. In turn, the bootstrap and iterated bootstrap provided intuitive algorithms for realizing in statistical practice the benefits of erudite asymptotic improvements.

Mathematical logicians investigating the notion of proof had greatly refined the concept of algorithm by mid-century (cf. Berlinski (2001)). Through the technological development of digital computers, programming languages, video displays, printers, and numerical linear algebra, stable computational algorithms enriched the statistician's toolbox. In consequence, a wider range of statistical procedures, numerical and graphical, became feasible. Case studies and experiments with artificial data offered non-probabilistic ways of understanding the performance of statistical procedures. The fundamental distinctions among data, probability model, pseudo-random numbers, and algorithm returned to prominence. The extent to which deterministic pseudo-random sequences can imitate properties of random variables received more attention (cf. Knuth (1969)). It became clear once again that data is not certifiably random. Computing technology provided a new environment in which to extend and reconsider statistical ideas developed with probability technology. The bootstrap is a case in point.

From our present technological standpoint, statistics is the development, study, and implementation of algorithms for data-analysis.

How is a data-analytic algorithm to be understood? One answer, offered by Brillinger and Tukey (1984), addresses the gap between statistical theory and data-analytic techniques:

If our techniques have no hypotheses, what then do they have? How is our understanding of their behavior to be described?

As a generalization of an umbra within a penumbra. Here there are at least three successively larger regions, namely:

1. An inner core of proven quality (usually quite unrealistically narrow) . . .
2. A middle-sized region of understanding, where we have a reasonable grasp of our technique's performance . . .
3. A third region, often much larger than the other two, in which the techniques will be used . . .

For example, the inner core of understanding could be an analysis under a simple probability model; the middle core could be asymptotic analyses and simulations under substantially more general probability models together with salient case studies; and the outer core would contain data analyses that use the techniques. In reality, data consists of scientific and other context as much as numerical observations.

For some statistical problems, such as classification of handwritten digits, probability models may not generate effective procedures. Breiman (2001) observed:

If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

His paper emphasized algorithmic modeling techniques that treat the data mechanism as essentially unknown.

How are data-analytic algorithms to be implemented? One answer, offered by the Omega-hat project (www.omegahat.org), focuses on open-source development of the next generation of statistical computing paradigms, environments, and software. The project provides an optionally typed language that extends both S and Java and a customizable, multi-threaded interpreter; and it encourages participation by those wanting to extend computing capabilities in one of the existing languages for statistical computing, by those interested in distributed or web-based statistical software, and by those interested in the design of new statistical languages.

This answer recognizes that software provides a powerful new medium for expressing statistical ideas. The Introduction to M. McLuhan's (1964) book *Understanding Media: The Extensions of Man* began:

In a culture like ours, long accustomed to splitting and dividing all things as a matter of control, it is sometimes a bit of a shock to be reminded, in operational and practical fact, that the medium is the message.

In other words, the nature of a medium has at least as much effect on human activity as does its content, which itself is just an older medium that is being expressed through the newer medium. In this manner, leading edge statistical computing environments stand to influence core ideas about statistics.

Fraser's paper examines pros and cons of conditional versus unconditional inference in classical probability models for data where the parameter of interest is one-dimensional. His examples indicate that these approaches can yield procedures with differing probabilistic properties. A diversity of answers is to be expected once we recognize the difference between data and probability model. In my discussion, I will consider: (a) the construction of simultaneous confidence sets, a problem that intrinsically has multiple answers with different properties; and (b) estimation of the means in a two-way layout with one observation per combination of factor levels, a typical multiparametric problem where neither sufficiency nor ancillarity is of immediate help. I will argue that, within the statistical environment created through technological advances in asymptotic theory and computing, the role of ancillarity and sufficiency is narrow.

Narrow is not the same as none. For example, Hájek's convolution theorem in locally asymptotically normal families has a form in which asymptotic sufficiency, asymptotic ancillarity, and Basu's theorem suggest a heuristic interpretation of necessary and sufficient conditions for consistency of parametric bootstrap distributions. The interested reader is referred to Beran (1996a) and, for a tangentially related nonparametric discussion, to van

Zwet and van Zwet (1999).

2. SIMULTANEOUS CONFIDENCE SETS

Coverage probability under a probability model does not, by itself, determine a confidence set. A further design goal, whether minimum expected geometrical size or equal conditional coverage probabilities, is needed to construct the confidence set. Geometrical size may be of interest if the confidence set is to serve as a set-valued estimator of the parameter. Equal conditional coverage probabilities may be of interest if the conditioning variable reflects a real feature in the data. An experienced statistician selecting a data-analytic algorithm will consider the context, and aims of the analysis as well as probability models.

The construction of simultaneous confidence sets has raised issues analogous to those in Fraser's second section. Consider a statistical model in which a sample X_n of size n has joint probability distribution $P_{\theta,n}$, where $\theta \in \Theta$ is unknown. The parameter space Θ is an open subset of a metric space, whether of finite or infinite dimension. Of interest is the parametric function $\tau = T(\theta)$, where T is a specified function on Θ . Suppose that τ has *components* $\{\tau_u = T_u(\theta): u \in U\}$, U being a metric space, which jointly determine τ . For each u , let C_u denote a confidence set for the component τ_u . By simultaneously asserting the confidence sets $\{C_u: u \in U\}$, we obtain a simultaneous confidence set C for the components $\{\tau_u\}$.

If the components $\{\tau_u\}$ are deemed logically similar, the statistician may wish to construct the confidence sets $\{C_u\}$ in such a way that

$$(2.1) \quad P_{\theta,n}[C_u \ni \tau_u] \text{ is the same } \forall u \in U$$

and

$$(2.2) \quad P_{\theta,n}[C_u \ni \tau_u, \forall u \in U] = P_{\theta,n}[C \ni \tau] = \beta.$$

Property (2.1) is called *balance*. It reflects the wish that the confidence set C treat the logically similar components τ_u in an even-handed way while controlling the simultaneous coverage probability (2.2). The balance constraint is a cousin to the equal conditional coverage probability condition treated in Fraser's second section.

One general approach starts with a *root* $R_{n,u} = R_{n,u}(X_n, \tau_u)$ for each component τ_u . The root may or may not be an exact pivot. Let \mathcal{T}_u and \mathcal{T} denote, respectively, the ranges of $\tau_u = T_u(\theta)$ and $\tau = T(\theta)$. Every point in \mathcal{T} can be written in the component form $t = \{t_u: u \in U\}$. The simultaneous confidence sets to be considered are

$$(2.3) \quad C = \{t \in \mathcal{T}: R_{n,u}(X_n, t_u) \leq c_u(\beta), \forall u \in U\}.$$

The technical problem is to devise critical values $\{c_u(\beta)\}$ so that, to a satisfactory approximation, C is balanced and has simultaneous coverage probability β for the $\{\tau_u\}$.

Let $H_{n,u}(\cdot, \theta)$ and $H_n(\cdot, \theta)$ denote the left-continuous cumulative distribution functions of $R_{n,u}$ and of $\sup_{u \in U} H_{n,u}(R_{n,u}, \theta)$ respectively. If θ were known and the two cdf's just defined were continuous in their first arguments, an oracle choice of critical values for the component confidence sets would be $c_u(\beta) = H_{n,u}^{-1}[H_n^{-1}(\beta, \theta), \theta]$. The oracle component confidence set

$$(2.4) \quad C_u = \{t_u \in \mathcal{T}_u: R_{n,u}(X_n, t_u) \leq c_u(\beta)\} = \{t_u \in \mathcal{T}_u: H_{n,u}(R_{n,u}, \theta) \leq H_n^{-1}(\beta, \theta)\}$$

has coverage probability $H_n^{-1}(\beta, \theta)$ for τ_u . The oracle simultaneous confidence set C , defined through (2.3), has coverage probability β for τ by definition of $H_n(\cdot, \theta)$. In historically influential special cases, this oracle construction can be carried out because neither $H_{n,u}$ nor H_n depends on the unknown θ .

Example 1. Suppose that X_n has a $N(A\gamma, \sigma^2 I_n)$ distribution, where the vector γ is $p \times 1$ and the matrix A has rank p . The unknown parameter $\theta = (\gamma, \sigma^2)$ is estimated by $\hat{\theta}_n = (\hat{\gamma}_n, \hat{\sigma}_n^2)$ from least squares theory. Suppose that the root

$$(2.5) \quad R_{n,u} = |u'(\hat{\gamma}_n - \gamma)| / \hat{\sigma}_{n,u},$$

where u is a p dimensional vector and $\hat{\sigma}_{n,u}^2 = u'(A'A)^{-1}u\hat{\sigma}_n^2$. The roots $\{R_{n,u}\}$ are identically distributed, each having a t distribution, folded over at the origin, with $n - p$ degrees of freedom.

Suppose that U be a subspace of R^p of dimension q . Then $\sup_{u \in U} R_{n,u}$ is a continuous pivot (cf. Miller 1966, Chap. 2, Sec. 2). In this instance, the oracle balanced simultaneous confidence set defined by (2.3) and (2.4) coincides with Scheffe's simultaneous confidence intervals for the linear combinations $\{u'\gamma: u \in U\}$.

Example 2. Specializing to a balanced one-way layout, suppose that U consists of all pairwise contrasts. The parameter γ is just the vector of means in this case of the linear model. Then $\sup_{u \in U} R_{n,u}$ is a continuous pivot (cf. Miller 1966, Chap. 2, Sec. 1). In this instance, the oracle balanced simultaneous confidence set defined by (2.3) and (2.4) coincides with Tukey's simultaneous confidence intervals for all pairwise differences in means.

The exact pivots used by Tukey and Scheffé in constructing their respective balanced simultaneous confidence intervals do not exist in most probability models. However, bootstrap techniques enable more general construction of simultaneous confidence sets that behave asymptotically like oracle simultaneous confidence sets. Suppose that $\hat{\theta}_n$ is a consistent estimator of θ . Replacing θ by $\hat{\theta}_n$ in the oracle critical values that appear in (2.4) yields bootstrap simultaneous confidence sets for the $\{\tau_u\}$. A Monte Carlo approximation to the bootstrap critical values requires only one round of bootstrap sampling. Computation of the supremum over U may require further approximations when the cardinality of U is not finite. In practice, the case of a finite number of components $\{\tau_u\}$ is both approachable and important. Theorem 4.1 in Beran (1988) provides sufficient conditions under which the

bootstrap simultaneous confidence set is asymptotically balanced and has asymptotic overall coverage probability β .

The balance condition (2.1) on the simultaneous confidence sets is a design element that can be modified at will. Technically speaking, we could seek specified proportions among the componentwise coverage probabilities. (I am not aware of a problem where this would be useful). The Tukey and Scheffé exact constructions and, more generally, the bootstrap construction are readily modified to handle this design goal. On the other hand, balance has not been found compelling in situations where the components $\{\tau_u\}$ are not logically comparable.

Example 3. Given an independent identically distributed sample from the $N(\mu, \sigma^2)$ distribution, it is easy to construct a balanced simultaneous confidence set of coverage probability β for the pair (μ, σ^2) . However, this is not a popular procedure, no doubt because the parameters μ and σ^2 are logically dissimilar.

The discussion in this section illustrates how advances in asymptotic and computer technology have given statisticians the ability to explore beyond the statistical principles of earlier eras, principles whose formulation captures, as in amber, the technological environment of their times.

3. MULTIPARAMETRIC TWO-WAY LAYOUTS

Consider a high-dimensional two-way layout with one observation per combination of factor levels. Factor k has p_k levels $\{t_{kj}: 1 \leq j \leq p_k\}$, which may be nominal or ordinal. Such a two-way layout is associated with experimental designs, grayscale images, and gene chips. Subscripting is arranged so that, for an ordinal factor, the factor levels are a strictly increasing function of subscript. A simple probability model asserts that

$$(3.1) \quad y_{ij} = m_{ij} + \epsilon_{ij}, \quad 1 \leq i \leq p_1, 1 \leq j \leq p_2,$$

where the $\{y_{ij}\}$ are the observations, $m_{ij} = \mu(t_{1i}, t_{2j})$, and the errors $\{\epsilon_{ij}\}$ are independent, identically distributed $N(0, \sigma^2)$ random variables. The function μ and the variance σ^2 are unknown. A basic problem is to estimate the means $\{m_{ij}\}$ and σ^2 .

For the means in model (3.1), the minimum variance unbiased (MVU) estimator and the minimum quadratic risk location equivariant estimator both coincide with the raw data. This estimator is unacceptable in contexts such as image processing or estimation of response surfaces. Indeed, Stein (1956) showed that the MVU is inadmissible under quadratic loss whenever the number of factor-level combinations $p = p_1 p_2$ exceeds 2. Neither reduction by sufficiency nor by ancillarity suggests a satisfactory estimator of the means in model (3.1). A partial exception to this claim holds for the one-way layout with nominal factor levels but does not handle ordinal factor levels (cf. Beran (1996b)).

What does work is regularization, the use of a constrained fit to the means that trades bias for variance so as to achieve lower risk in estimating the means of the two-way layout. Regularization is an estimation strategy for models that have many or an infinite number of unknown parameters—models that play a prominent role in modern statistics. A regularized fit is typically constructed in three stages. First, we devise a candidate class of constrained mean estimators that individually express competing prior notions about the unknown means. Secondly, we estimate the risk of each candidate estimator under a general model that does *not* assume any of the prior notions in step one. Thirdly, we define the regularized fit to be a candidate fit that minimizes estimated risk or a related criterion. This regularized fit may be interpreted as the trend discernible in the noisy observations.

In the two-way layout, let y denote the $p \times 1$ vector obtained by ordering the observations $\{y_{ij}\}$ in mirror dictionary order: the first subscript runs faster than the second subscript. Let m denote the similarly vectorized means $\{m_{ij}\}$. Model (3.1) asserts that the distribution of y is $N(m, \sigma^2 I_p)$. For $k = 1, 2$, define the $p_k \times 1$ vector u_k and the $p_k \times p_k$ matrices J_k, H_k by

$$(3.2) \quad u_k = p_k^{-1/2}(1, 1, \dots, 1)', \quad J_k = u_k u_k', \quad H_k = I_{p_k} - J_k.$$

For each k , the symmetric idempotent matrices J_k and H_k have rank (or trace) 1 and $p_k - 1$ respectively. They are thus orthogonal projections that decompose R^{p_k} into two mutually orthogonal subspaces of dimensions 1 and $p_k - 1$. The identity $I_{p_k} = J_k + H_k$ implies that

$$(3.3) \quad m = (I_{p_2} \otimes I_{p_1})m = P_0 m + P_1 m + P_2 m + P_{12} m,$$

where $P_0 = J_2 \otimes J_1$, $P_1 = J_2 \otimes H_1$, $P_2 = H_2 \otimes J_1$, and $P_{12} = H_2 \otimes H_1$. Equation (3.3) gives, in projection form, the ANOVA decomposition of a complete two-way layout of means into overall mean, main effects, and interactions.

Certain penalized least squares criteria generate a class of candidate estimators by restricting, in varying degree, the ANOVA decomposition. Let A_k be any matrix with p_k columns such that $A_k u_k = 0$. Examples of such *annihilator* matrices are $A_k = H_k$, suitable when factor k is nominal, and A_k equal to the d -th difference matrix, suitable when factor k is ordinal with equally spaced factor levels $\{t_{kj}\}$. Let $B_k = A_k' A_k$ and define $Q_1 = J_2 \otimes B_1$, $Q_2 = B_2 \otimes J_1$, and $Q_{12} = B_2 \otimes B_1$. Let $A = \{A_1, A_2\}$ and let $\nu = (\nu_1, \nu_2, \nu_{12})$ be any vector in $[0, \infty]^3$. The *candidate penalized least squares* (PLS) estimator of m is $\hat{m}_{PLS}(\nu, A) = \operatorname{argmin}_m S(m, \nu, A)$, where

$$(3.4) \quad S(m, \nu, A) = |y - m|^2 + m'(\nu_1 Q_1 + \nu_2 Q_2 + \nu_{12} Q_{12})m.$$

The symmetric matrix B_k has spectral decomposition $U_k \Lambda_k U_k'$, where $\Lambda = \operatorname{diag}\{\lambda_{ki}\}$ is diagonal with $0 = \lambda_{k1} \leq \lambda_{k2} \leq \dots \leq \lambda_{kp_k}$ and the eigenvector matrix U_k is orthonormal with

first column equal to u_k . Let $f_{ij}(\nu) = [1 + \nu_1 \lambda_{1i} e_{2j} + \nu_2 e_{1i} \lambda_{2j} + \nu_{12} \lambda_{1i} \lambda_{2j}]^{-1}$, where $e_{k1} = 1$ and all other $\{e_{kj}\}$ vanish. Vectorize the $\{f_{ij}(\nu)\}$ in mirror dictionary order to obtain the vector $f(\nu)$ and let $z = (U_2 \otimes U_1)'y$. It follows readily that the candidate PLS estimator is the shrinkage estimator

$$(3.5) \quad \hat{m}_{PLS}(\nu, A) = (U_2 \otimes U_1) \text{diag}\{f(\nu)\}z.$$

Let $\xi = (U_2 \otimes U_1)'m$ and let $\text{ave}(h)$ denote the average of the components of vector h . The normalized quadratic risk of the (usually biased) candidate estimator (3.5) is

$$(3.6) \quad p^{-1} \mathbb{E} |\hat{m}_{PLS}(\nu, A) - m|^2 = \text{ave}[f^2(\nu)\sigma^2 + (1 - f(\nu))^2 \xi^2].$$

The operations inside the average are performed componentwise, as in the S language.

Having devised a variance estimator $\hat{\sigma}^2$ by some form of pooling, say, we may estimate the risk (3.6) by

$$(3.7) \quad \hat{r}(A, \nu) = \text{ave}[f^2(\nu)\hat{\sigma}^2 + (1 - f(\nu))^2(z^2 - \hat{\sigma}^2)].$$

This is just Stein's unbiased risk estimator with σ^2 replaced by $\hat{\sigma}^2$. For a specified class \mathcal{A} of annihilator pairs A , we define the *adaptive PLS* estimator of m to be the candidate PLS estimator with smallest estimated risk:

$$(3.8) \quad \hat{m}_{PLS} = \hat{m}_{PLS}(\hat{\nu}, \hat{A}), \quad \text{where } (\hat{\nu}, \hat{A}) = \underset{(A, \nu) \in \mathcal{A} \times [0, \infty]^3}{\text{argmin}} \hat{r}(A, \nu).$$

This adaptive estimator is an empirical approximation to the oracle candidate PLS estimator that minimizes the unknown risk (3.6) over $(A, \nu) \in \mathcal{A} \times [0, \infty]^3$.

Computational algorithms, case studies, and multiparametric asymptotics for \hat{m}_{PLS} are developed in Beran (2002). Under model (3.1), subject to restrictions on the richness of the annihilator class \mathcal{A} and to assumptions that ensure consistency of $\hat{\sigma}^2$, the risk of the adaptive PLS estimator \hat{m}_{PLS} converges to that of the oracle candidate estimator as the number of factor-level combinations tends to infinity. By construction, this limiting risk cannot exceed that of the MVU estimator. In case studies, it is not unusual for the adaptive PLS estimator to reduce risk by a factor of three or more over that of the MVU estimator. For two-way layouts with nominal factors, the adaptive PLS estimator generated by $A_k = H_k$ essentially coincides with the multiple shrinkage estimator studied by Stein (1966). For two-way layouts with ordinal factors, the adaptive PLS estimator based on local polynomial annihilators can be strikingly more efficient than the MVU estimator and is akin to spline fits in two-way functional data-analysis.

The foregoing discussion of the two-way layout illustrates a technology developed over the past five decades for better estimation in multiparametric and nonparametric models. The role of sufficiency and ancillarity has been inconsequential in this substantial portion of modern statistics.

REFERENCES

- Beran, R. (1988). Balanced simultaneous confidence sets. *J. Amer. Statist. Assoc.* **83** 679–686.
- Beran, R. (1996a). Diagnosing bootstrap success. *Ann. Inst. Statist. Math.* **49** 1–24.
- Beran, R. (1996b). Stein estimation in high dimensions: a retrospective. In *Madan Puri Festschrift* (E. Brunner and M. Denker, eds.) 91–110. VSP, Zeist.
- Beran, R. (2002). ASP algorithms for denoising two-way layouts. Preprint available at www.stat.ucdavis.edu/~beran/two.pdf.
- Berlinski, D. (2001). *The Advent of the Algorithm*. Harcourt, New York.
- Breiman, L. (2001). Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statist. Sci.* **16** 3, 199–231.
- Brillinger, D. R. and Tukey, J. W. (1984). Spectrum analysis in the presence of noise: some issues and examples. In *The Collected Works of John Tukey II* (D. R. Brillinger, ed.) 1001–1141. Wadsworth, Monterey, CA.
- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7** 1–26.
- Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. Hafner, New York.
- Knuth, D. E. (1969). *The Art of Computer Programming, Vol 2: Seminumerical Algorithms*. Addison Wesley, Reading MA.
- McLuhan, M. (1964). *Understanding Media: The Extensions of Man*. McGraw Hill, New York.
- Miller, R. (1966). *Simultaneous Statistical Inference*. McGraw-Hill, New York.
- Savage, L. J. (1954). *The Foundations of Statistics*. Wiley, New York.
- Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. Third Berkeley Symp. Math. Statist. Probab.* (J. Neyman, ed.) 197–206. Univ. California Press, Berkeley.
- Stein, C. (1966). An approach to the recovery of inter-block information in balanced incomplete block designs. In *Festschrift for Jerzy Neyman* (F. N. David, ed.) 351–364. Wiley, New York.
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading MA.
- van Zwet, E. W. and van Zwet, W. R. (1999). A remark on consistent estimation. *Math. Methods Statist.* **8** 277–284.
- Wald, A. (1950). *Statistical Decision Functions*. Wiley, New York.