Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

DAVID R. BRILLINGER

Statistical Inference for Random Processes*

1. Introduction

Statistics is concerned with data collection, data analysis, data reduction, data modelling and inference. Its primitive concept is that of data. Statistics is part of the methodology of science — pure and applied. It is pertinent to the various goals of science proper: explanation and understanding, prediction and control, discovery and application, justification-classification. Two things at the heart of science are observation and inference. Inference may be deductive, arguing from the premises to conclusions, or what is the major process in science, inductive, intuiting from the specific to the more general.

Statistical inference is concerned with making statements that go beyond the data collected. Its traditional paradigm is that of from the sample to the population or parameter. The strength of statements made depends on the situation at hand. There are several schools of statistical inference. The schools are often in conflict; however, these days, their chosen principles are fairly clear.

By now statistics has amassed quite a collection of procedures for drawing inferences from data; however, with the passage of time, the data of concern has gotten steadily more complex. This essay is concerned with statistical inference in general and for random process data in particular. In barest detail a random process is an indexed family of random variables (or chance quantities). In operational use a random process is a random function, or random measure, or random generalized function with domain that is temporal or spatial or spatial–temporal. Its values have coordinates. Its realizations are: curves, surfaces, shapes, figures,

* Prepared with the partial support of the National Science Foundation, Grant CEE-7901642 and while the author was a Guggenheim Fellow.

sequences and the like. It relates to situations where things move and change.

We begin with an example of statistical inference for random processes taken from our own experience. The example is one with a precise experimental setup yet, apparently, inferences may not be drawn from direct examination of the data or after the realization of new experiments. Rather, a statistical concept of some subtlety is required to unravel the situation. We remark that the statistician is concerned with the probabilistic conceptualization of natural processes. At the same time he is a guardian of a collection of tools that bring order to complex data sets, tools which have had real successes. The remaining sections of the paper reflect these two aspects. Scientific investigation and modelling are discussed in general terms. Process data analysis and its aims are discussed in particular terms.

Though it is not brought out specifically in the paper, mathematics is always present for the statistician. Sometimes, especially in the theory of random processes, his work is indistinguishable from mathematics. At other times mathematics is a potent heuristic aid for planning data collection and analyzing data at hand.

2. An example

A sequence of nerve impulses, or spike train, is a common form of neurophysiological data. The times of the pulses correspond to the times at which a particular neuron fires off. The heights of the pulses are nearly constant and, provided the experimental conditions are reasonably fixed and the experiment is not continued too long, the character of the spike train is not seen to be evolving with time. It appears that this kind of data may be reasonably modelled as a piece of a realization of a stationary point process on the real line. Such a process may be defined as a random process whose realizations $N(\cdot)$ are non-negative integer-valued Borel measures on R with the (stationarity) property that the probability that $N(I_1+t) = n_1, \ldots, N(I_K+t) = n_K$ does not depend on t for I_K a Borel subset of R and $K = 1, 2, \ldots$ Suppose that the observed times of consecutive pulses, for a given spike train, are $t_1, \ldots, t_n$. Then a key role is played in the example by the empirical Fourier transform

$$d(\lambda) = \sum_{j=1}^{n} \exp\{-i\lambda t_j\} = \int_T \exp\{-i\lambda t\} N(dt),$$

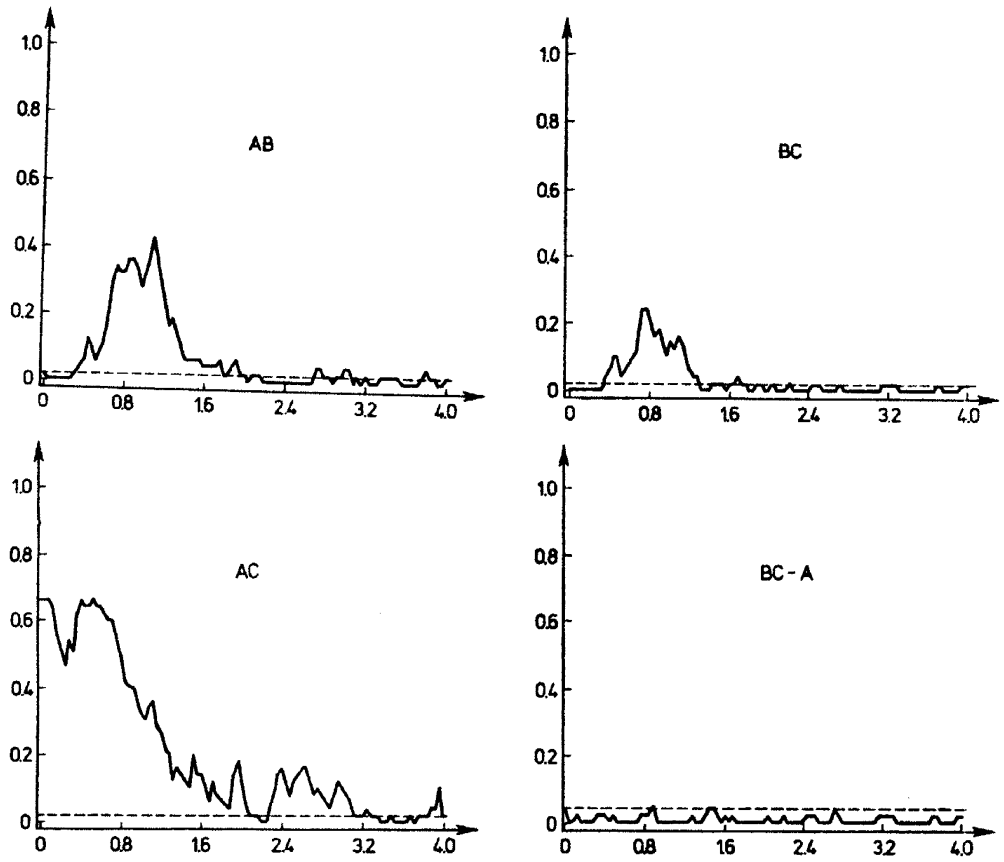
where $\lambda \in R$ and T is the observation domain.

In our example, spike trains could be recorded simultaneously for three neurons A, B, C of *Aplysia californica*. It was "known" that neuron A was driving neurons B and C. It was not known whether there was some separate connection between neurons B and C and this was the scientific question of interest. (Details may be found in Brillinger *et al.*, *Biol. Cybernetics* **22** (1976), 213–228.)

A useful statistic for measuring the degree of association (at frequency λ) of two empirical spike trains, A and B, is the sample coherency

$$\hat{R}_{AB}(\lambda) = \hat{f}_{AB}(\lambda) / \sqrt{\hat{f}_{AA}(\lambda)\hat{f}_{BB}(\lambda)}$$

where $\hat{f}_{AB}(\lambda)$ is obtained by averaging values of $d_A(\mu)\overline{d_B(\mu)}$ for μ in a neighborhood of λ. Provided the same averaging is employed in forming $\hat{f}_{AA}(\lambda)$, $\hat{f}_{BB}(\lambda)$ one has $|\hat{R}_{AB}(\lambda)|^2 \leqslant 1$, with values near 1 corresponding to strong association. The Figure shows the functions $|\hat{R}_{AB}|^2$, $|\hat{R}_{BC}|^2$, $|\hat{R}_{AC}|^2$ for one particular set of experimental data and the spike trains are indeed "found".



to be associated in pairs. The issue is whether the association of neuron B and C results totally from their both being driven by neuron A, o whether they have some association (connection) beyond that. To the ex tent that relationships involved are well enough captured by quadratic stat istics, one can address such questions by partial coherencies.

The sample partial coherency of trains B and C given train A is

$$\hat{R}_{BC\cdot A} = (\hat{R}_{BC} - \hat{R}_{BA}\hat{R}_{AC}) / \sqrt{(1-|\hat{R}_{BA}|^2)(1-|\hat{R}_{CA}|^2)}.$$

One has $|\hat{R}_{BC\cdot A}|^2 \leqslant 1$, with values near 0 corresponding to weak associ ation of trains B and C having "removed" the effects of train A. Th Figure presents this function for the given data. There is the stronges suggestion of no direct connection between neurons B and C.

To formalize this "strongest suggestion" the 5 per cent significanc line is given in each plot, as the horizontal dashed line. Were there n separate connection of B and C, the probability of this line being exceede at a given frequency would be (approximately) 0.05.

The situation now reached is typical of what happens in science an what statistical inference has to offer. The hypothesis (of no direct connec tion) cannot be verified absolutely; hence it is given an opportunity t show itself false. What has happened is that the data have shown them selves compatible with the hypothesis up to the limits of the inherent vari ation present. Probability has been used to formalize this last.

3. Scientific investigation

In an earlier paper on our topic, (*J. Royal Statistical Society* **A 130** (1967) pp. 457–477) M. S. Bartlett sets up a "ladder diagram" of scientific en quiry of the following form:

(Theory)		(Practice)
model	↔	planning/design
deduction	→	data collection
induction	←	data analysis
new model	↔	new planning/design

Things are initiated by some idea, question or problem. Then one move down and across the steps as work progresses. (Similar schemata hav been given by G. E. P. Box, *J. American Statistical Association* **71** (1976 pp. 791–799, and H. Mohr, *Structure and Significance of Science*, Springe Verlag (1977).) Deductions from the model play a broad role and a narro

one. Broadly they may be predictions that science and technology use to make progress. Narrowly, they may be used just to validate the model with extant data. (Statisticians have been much concerned with this last.)

An essential feature of the whole investigative procedure is its cyclic/iterative character: ... deduction to induction to deduction to ...

4. Process data

Commonly the term process has referred to a phenomenon which showed a continuous change with time. However, the idea has been substantially abstracted with the time parameter allowed to be discrete, multidimensional, set-valued and function-valued amongst other things. Further, any requirement of continuity has been directly adapted to the situation at hand.

Process data refers to information that has been derived by observation of the process at some collection of "time" values. The information will often have numerical form; however, its values can lie in some general structured space. We shall write process data as $\{Y(t), t \in T\}$, T denoting the observation domain.

In using the term we have in mind things like: the recorded arrival times of individual photons collected by a telescope aimed in some direction, stereoscopic photographs from a distance of some land or sea surface, the collection of time series recorded at an array of sensors after a pulse of energy is input to the earth, measurements of X-ray absorption by the head as a function of the direction of a submitted X-ray beam, the distribution of earthquakes through space and time. In discussions of process data it is usual to work in situations for which the number of realizations, n, of the process $Y(t)$ is much less than the dimension, p, of the observation domain T. Multivariate data analysis, in contrast, concentrates on the case $n \gg p$.

Thanks to the dramatic advances in equipment and instrumentation during the past 30 years, researchers have effective tools for dealing with the collection of many sorts of process data, e.g. ultrafast phenomena and spatial-temporal fields. Issues arising are: data selection (auxiliary variates?), data storage (device, structure), data retrieval, data display, data auditing and flagging. Particular aspects of the process of interest affecting how this is done are: data type, data frequency content/dynamic range/information content and whether one is working in real-time

or off-line. It is clear that digital computers are important. Optical computers are now beginning to play an important role as well, e.g. in data smoothing and Fourier transforming.

As indicated in the ladder diagram, the model, deductions from the model and the design of the investigation affect data collection. We shall return to these stages later.

5. Aims of process data analysis

A time series, $Y(t)$, is a particular type of process for which t and $Y(t)$ are real-valued. J. W. Tukey, *Directions in Time Series* (Eds. D. R. Brillinger and G. C. Tiao), Institute of Mathematical Statistics (1980), has listed the following aims of time series analysis:

1. discovery of phenomena,
2. modelling,
3. preparation for further inquiry,
4. reaching conclusions in statistical terms,
5. assessment of predictability,
6. description of variability.

These apply to the general process case as well. Having in mind the great variety of process data, we may also mention: control, classification, establishing causation, description of relationship, summarization, removal of concommitant variation, measuring degree of association, signal reconstruction and enhancement, questioning conformity of theory to data, focusing information, precise measurement of constants, comparative analysis.

The neurophysiological example that we presented earlier was concerned with reaching conclusions; however, the technique employed, Fourier analysis, is well-suited to discovering unsuspected phenomena.

We have available today a broad collection of methods for meeting the aims above. Various factors enter into the choice of method for an intended analysis. One of the most important is the degree of urgency involved in the situation at hand. A second is the computing facilities available.

6. Methods for process data analysis

At the operational level the methods available for process data analysis depend upon the type of process of concern; however, there do exist a number of techniques of quite broad applicability. We shall concentrate

on these. Further any technique employed will depend intimately on the aim of the analysis.

Manipulations possible for process data depend upon the particular character of the process under study as well as the computational and instrumental facilities available. Linear forms in the data are by far the most common. They may be real-valued or function-valued. Included are Fourier and other transforms and least squares projections. In many cases they are chosen to have high information content.

It is clear that one can contemplate working with quadratic and other polynomial forms in the data. This has proved to be successful on many occasions. Great advantages of such forms are that they may be manipulated directly and that computational devices for their evaluation are often available.

The step away from polynomial forms is a long one. Experience and insight have sometimes suggested particular statistics to work with Alternatively, models of the situation of concern have proved a rich source We will return to the concept of model shortly.

Things computed and displayed are located at several levels.Some things are the primary goals of the work. Other things are intended to indicate the uncertainty (or instability) of those primaries. Yet other quantities are evaluated to examine and challenge assumptions (the model) that drove the analysis.

Among specific methods applicable to process data are: spectrum analysis, smoothing, inversion, likelihood, Kalman–Bucy, clustering, re-expression, dimensional reduction, contingency, analysis of variance, least squares, simulation. Specific algorithms exist for their application to many types of data. However, there are continual difficulties that arise in practice and complicate the use of the algorithms. These include: missing data, out-of-line data values, measurement error, concomitant variation, extra structure in the data, artifacts, heterogeneous data, censored data, biased collection procedure, jitter, discretization error. A broad variety of procedures now exist for dealing with these difficulties.

7. One important method

In a surprisingly large number of situations, the Fourier transform provides a meaningful method for handling process data. It is broadly defined, flexible and has useful mathematical, statistical and computational properties. We have already indicated the form of the Fourier transform

of some point process data. If instead we had planar data on a continu ous process, it would take the form

$$d(\lambda_1, \lambda_2) = \iint_T Y(t_1, t_2)\exp\{-i(\lambda_1 t_1 + \lambda_2 t_2)\}dt_1 dt_2,$$

T denoting the domain of observation. In many situations it turns out to b helpful, and sometimes even crucial, to insert a convergence factor, ψ forming for example

$$\iint \psi(t_1, t_2)\, Y(t_1, t_2)\exp\{-i(\lambda_1 t_1 + \lambda_2 t_2)\}dt_1 dt_2,$$

the support of ψ being contained in T, ψ being approximately 1, but taper ing off to 0 as it approaches the boundary of T. The last expression ex tends quite directly to the case of a generalized (Schwartz–Bruhat) proces over an abelian locally compact group.

It should be no surprise that the Fourier transform of process data i useful for handling convolutional relationships. (Indeed, this was one rea son for its use in the example of Section 2.) It is also useful for examinin a process for phenomena at "frequency" λ. One way this is done is vi the periodogram, $|d(\lambda)|^2$, or some smoothed form of this last. The fiel of seismology provides two pertinent examples. Consider the suite of tim series recorded by an array of seismometers. Following an earthquake seismic signal may move across the array. A periodogram type analysi of this data can be used to estimate the direction of the source of the seis mic energy and the velocity with which it is travelling (and this may b done for individual temporal frequency bands). By doing this analysi for successive time periods, changes in the energy source may be note and associated phenomena viewed. Aki and Chouet, *J. Geophysics Res.* **8** (1975), pp. 3322–3342, provide an example wherein, following an explosion Fourier analysis first shows energy coming from the appropriate directio with the expected velocity, this is then followed by energy arriving fro all directions with various velocities — apparently the result of back scattering. Bolt *et al.*, *Earthquake Engineering and Struct. Dynam.* **10** (1982) pp. 561–573, provide another example of this sort of analysis. In thei case, records from a nearby earthquake were processed. The apparen direction of the source of seismic energy was seen to shift with time. Thi may have been the first experimental measurement of a seismic dislocatio moving along a rupturing fault. In each case, Fourier analysis allowed on to "discover" the presence of suspected scientific phenomena.

One tremendous statistical advantage of employing Fourier analysi

is that, in the case of a stationary process, the problem is turned into one involving independent identically distributed random variates.

8. Modelling

An ubiquitous concept in the work of statisticians (and indeed of all researchers) is that of model. A variety of meanings are attached to the word. (Some of these are reviewed by P. Suppes, *Synthese* **12** (1960) pp. 287–301.) It is often taken to mean a theory. With a model at hand, much of a researcher's work becomes deductive and manipulative. The greatest difficulties lie in creating pertinent models. Statisticians end up with a schizophrenic attitude to them. This is well illustrated by two statements of G. E. P. Box: "Statistics is or should be the art and science of building scientific models which (necessarily) involve probability.", "Since all models are wrong the scientist cannot obtain a 'correct' one by excessive elaboration."

Workers have developed a number of methods for assessing, impartially, the strength of evidence for or against a particular model, (i.e. for model validation) and for estimating the values of quantities characterizing a given model (parameters). Much work with models is concerned with investigating them theoretically and examining their goodness-of-fit empirically.

The vast majority of statistical analyses rest on a probability model of a process under investigation. Consideration of a random entity allows all of probability theory to be brought to bear on problems—in particular, for example, results concerning special random processes. In the case of a system (that is, a structure consisting of possible inputs, an operation and corresponding outputs) there now exists an immense literature concerning identification given data consisting of pieces of (process) input and corresponding pieces of (process) output. An essential practical distinction arises between situations in which the scientist can select (some of) the inputs and those where they are outside his control. Another distinction is whether the model is mechanistic (based on specific description of the natural components involved) or empirical (based on regularities that caught the researchers eye). The former is the fundamental one.

9. Statistical inference

A statistical inference is a map from data to an uncertain conclusion. The logic involved is multi-valued. The procedure is inductive. Statements made are correct only in some average sense. The statistician usually pro-

ceeds by building a chance model for the situation. Questions that aris include: is the constructed model adequate for the data? how shoul subjective information be incorporated? in what form should the con clusions be stated and what is then their meaning? are there importan unmeasured variables? what is the goal of the work? on what shoul probability statements be conditional? how is a better model to be dis covered? what parallel models should be considered? how should fact o preliminary analysis be incorporated? how should costs be inclu ded?

Uncertain conclusions drawn after a data analysis have various form and levels. At one extreme one has what Mosteller and Tukey, *Data Analy sis and Regression*, Addison–Wesley (1977), call a "concealed inference' wherein the data are so strong that no formalism or arithmetic are re quired to come to a solid conclusion. Indeed, the very goal of experimen tation is to end up with such certain conclusions. At another extreme a conclusion involves but an elementary indication of the suspected vari ability (stability) of some primary entity derived from the data at hand In between one has a broad collection of inference forms and tools. W mention: tests of significance, confidence regions, likelihood graphs, pos terior distributions, tolerance regions, standard errors, distance measures prob-values, fiducial probabilities, sensitivity analyses, simulations.

One of the major contemporary works on statistical inference for randon processes is that of U. Grenander, *Abstract Inference*, J. Wiley (1981) It is worth indicating some of the distinctions he recognizes and problem and procedures that he highlights. By his choice of the term "abstrac inference" he deliberately leaves ambiguous whether he means the sampl space (set of possible observations) or parameter space (values for quan tities characterizing the probability distribution at hand) or both to b "abstract". In the work he discusses each case. For inference he employs linear methods, likelihood based estimates and direct methods (the latte being based on common sense estimates). Classical statistical inferenc falls from the first two, once the appropriate structure is set up. To dea with the fact that classical procedures sometimes fail if the paramete space is too large, Grenander introduces the "method of sieves"—em ploying the classical procedure over a subset of the parameter space. Th method is like Tihonov regularization and, for example, leads to spline in the case of nonparametric regression. Related circles of ideas include penalized maximum likelihood, Courant regularization, Bayesian estima tion, ridge regression, and Stein estimates.

In the analysis of process data three situations requiring different

statistical techniques, occur in practice: the signal-like situation, the noise-like situation and the mixture of signal and noise situation. In the signal-like case records for the same circumstances differ chiefly by measurement noise, e.g. images under the same conditions, identical utterances by one individual. In the noise-like case realizations have quite different appearances, e.g. the roughness of two pieces of road surface, turbulent fields generated in repetitions of an experiment. The third case is a hybrid, e.g. an earthquake recorded near a sea storm. In the signal-like case interest often is to estimate the signal. Smoothing or deconvolution operations, including regularization, may be invoked. In the noise-like case interest lies in the population from which the realization came and, for example, what may be sought is a description of the variability present or of other underlying characteristics. Difficulties arise if one uses a technique developed for one case, with another. Comparison of signals requires generalization of classical ANOVA.

So-called inverse problems fall into the signal-like case. These include the problems of computerized tomography, image reconstruction and earth modelling. They may often be formulated as: $y = X\theta + \varepsilon$, with y, θ, ε lying in abstract spaces, with X a known operator and with y also given. The problem is to estimate the signal θ. Difficulties arise because of the presence of the noise ε and because X is often unbounded. The Tihonov regularization approach chooses as estimate the value of θ minimizing $\|y - X\theta\|^2 + \alpha\|\theta\|_0^2$ for some scalar α and θ lying in some normed space. In a number of cases the estimate may be written $\theta = (X'X + \alpha A)^{-1}X'y$, for A an operator.

Photon correlation spectroscopy provides an example of a noise case where one is interested in describing the variability present. In one application, similar particles suspended in a liquid are in motion with differing velocities. It is desired to estimate the distribution of velocities. To do this, the liquid is illuminated by a laser beam. The motion of the particles induces Doppler shifts of the laser frequency, specifically the autocovariance function of the scattered light is proportional to $1 + a|b(u)|^2$ at lag u where a is a constant and $b(u) = \int [(\sin uqv)/uqv] f(v)\,dv$, $f(v)$ being the desired velocity distribution and q a known constant. The autocovariance may be estimated from a photo-multiplier record of the fluctuating light. The function $f(v)$ may be estimated by regularizaton. One reference is Frost and Cummins, *Science* **212** (1981), pp. 1520–1522. They measure sperm motility.

It seems fair to say that once a stochastic model has been set down much of the work of statistical inference proceeds in a regular manner.

The book by I. V. Basawa and B.L.S. Prakasa Rao, *Statistical Inferen for Stochastic Processes*, Academic Press (1980) contains many results f a broad array of random processes. Difficulties arise on two fronts. Firs many of the results are based on approximations, so they need study i any particular situation. Second, and more importantly, there is the pro lem of obtaining a reasonable model. In seeking a model the researche typically turns to substantive theory and exploratory data analys (using J. W. Tukey's term). At some point the researcher has to have a insight. This is a subconcious act and there is little likelihood that it ca ever be made mechanical, but with today's marvellous visual displa devices and growing collection of exploratory data tools, environment for insight can be set. Process data typically involves an element of chang or movement, making visual displays especially appropriate.

10. Planning and experimental design

We conclude with a few comments on planning/design issues for proces data. The distinction between experimental and observational data i crucial. (In the system case—the distinction between chosen and natura input.) The quality of inferences that may be drawn depends dramaticall on which type of data is at hand. With observational data one has alway to be concerned that some unsuspected or "hidden" variable was control ling the situation, not the variables that showed themselves. Through th choice of factors to vary, through the design of input, through the use o randomization a researcher can validate his statistical inferences and mak efficient use of resources.

Once again many situations may be studied via the model $y = X\theta + \varepsilon$ provided one is flexible in definitions. Taking X such that $X'X$ is the iden tity has long been known to be an effective plan in elementary experi mental design. In the case of a process system, this leads to taking as inpu things like: Gaussian white noise, a homogeneous Poisson, pseudorandom binary noise and a train of chirp signals. A noteworthy phenomenon i that stimuli developed for experiments in one substantive field find us in other substantive fields. We mention the chirp signal moving from radar to exploration seismology, the sinusoid moving from power engin eering to laser spectroscopy, white noise moving from mechanical engin eering to nuclear magnetic resonance spectroscopy. An additional benefi of employing random stimuli is that hidden variables are neutralized as in traditional statistical experiments.

In the case of a nonlinear system, only a few input processes have beer

studied extensively. N. Wiener argued for the use of Gaussian white noise in the case of polynomial systems. It has led to satisfactory results in a number of physical situations.

11. Epilogue

Taking note of the site of this Congress *and* the site of the next, it would be remiss not to make specific mention of Jerzy Neyman. His following words are as true today as they were some twenty years ago: "Currently in the period of dynamic indeterminism in science, there is hardly a serious piece of research which, if treated realistically, does not involve operations on stochastic processes. The time has arrived for the theory of stochastic processes to become an item of usual equipment of every applied statistician." *J. Amer. Statist. Assoc.* **55** (1960), pp. 625–639.

THE UNIVERSITY OF CALIFORNIA
BERKELEY; U.S.A.