The Annals of Statistics
2002, Vol. 30, No. 6, 1619–1628

JOHN W. TUKEY AS "PHILOSOPHER"

BY A. P. DEMPSTER

Harvard University

Although not a traditional philosopher, John Tukey contributed much to our understanding of statistical science and empirical science more broadly. The former is represented by the light he shed on the relation of drawing conclusions to making decisions, and of how simple concepts like significance and confidence serve to back up or "confirm" empirical findings. Less successfully, he attempted inconclusively to sort out the ambiguities of R. A. Fisher's fiducial argument. His main effort, however, went to creating "exploratory data analysis" or EDA as a subfield of statistics with much to offer to ongoing developments in data mining and data visualization.

1. Introduction. John Tukey was a major player who projected unique and important messages across statistical science for more than four decades following World War II. His most visible genius was mathematical, but science was in his psyche, and his wartime experience as a statistical consultant and problem-solver led him to focus on using his talents to advance his chosen field of statistics. His belief in the crucial role of mathematics was strong, but he chose to exercise his talents not by developing and proving theorems, but rather by seeking ways to bridge the gap between mathematics and science. Work of this kind is almost certain to lead one to being labeled a "philosopher" by colleagues working in more conventionally constrained specialties.

In the foreword (hereafter JWT86) to Volumes III and IV of the *Collected Works* John asserts:

> The philosophy that appears in these two volumes is far more based on a "bottom up" approach than on a "top down" one,

meaning that whatever might be construed as philosophy in his writing was directly motivated by trying to get specific data analysis tasks right, not by trying to define and promote an overall framework for doing statistical science. So perhaps it is not philosophy at all, just mainstream statistics.

John's importance springs from a remarkable intellect, characterized by an ability to easily grasp complex details that more ordinary mortals are wont to see dimly with great difficulty. By repeatedly showing people how to think outside narrow boxes defined by most teachers and colleagues, he opened eyes to unrecognized possibilities. As a result he was highly valued as a consultant. His

Received September 2001; revised February 2002.
AMS 2000 subject classifications. 62A01, 62-07, 62F03, 01A70.
Key words and phrases. Conclusions, decisions, confirmatory, exploratory, Fisher, Neyman, Bayes.

discursive writing style could be infuriating, and his dominating personality could be intimidating, but he learned to project a tolerant image, and was generous of his time in the service of solving interesting problems, whether from lowly students or at high levels of a national establishment. Those of us lucky enough to be directly touched by him owe an especially deep debt of gratitude.

A few brief remarks can do little more than scratch the surface of the 1000 pages of the "philosophy" volumes, written over three plus decades largely in response to invitations from editors and conference organizers. Much of the material is repetitious, yet all is insightful and written with deliberate care. There is a wealth of statistical wisdom in these many pages, that should long fascinate both practitioners and historians seeking a snapshot focused mainly on the third quarter of the 20th century.

I make no claim that what follows is a thorough evaluation against the background of developments that John was reacting to as his career unfolded. The developments of this period were an outgrowth of fundamental sampling theories pioneered in the 1920s and 1930s, first by Fisher and subsequently by the frequentist school surrounding Jerzy Neyman. The period also saw the countervailing advocacy of Bayesian analysis beginning in the 1950s. John assessed many of these developments, as I shall sketch below. Recent decades have been pushed and pulled more by the continuing revolution in computing technologies than by conceptual change, whence, until new fundamental paradigms take root, the issues under discussion in the *Collected Works* retain contemporary relevance.

2. CDA and EDA. John's trademark emphasis on data analysis developed from his wartime work, especially with older colleague Charlie Winsor. John maintained that although formal study and Ph.D. training are needed, effective practice of statistics requires subsequent years of direct experience with applications. In the 1940s, John made pathbreaking contributions, most notably to spectral analysis of time series, and to procedures for multiple comparisons, emphasizing inferential methods based on sampling models that were the norm at the time. He later called this approach "confirmatory data analysis" to contrast it with the nonprobabilistic "exploratory data analysis" concept that became his major focus. His perceptions of the problems with applying sampling models, that ultimately led to his shifting emphasis from CDA to EDA, are well captured by Lyle Jones in introductory remarks to Volumes III and IV of the *Collected Works*:

> A theme emphasized in these volumes is that statistics is an empirically based discipline and that statisticians must be prepared for surprises in data. Almost invariably, when closely inspected, data are found to violate standard assumptions that are required in the application of classical methods of statistical inference. Starting from the premise that it is better to be "approximately right" rather than "exactly wrong" Tukey searches for procedures that are resistant to common violations of those assumptions ... Tukey is not content to adopt general purpose nonparametric procedures; he seeks solutions that aid in drawing meaningful conclusions and suggests procedures that give up little in

> efficiency. For example, by attending to distributions of residuals to a fit of a statistical model to the data, an improved fit may result from a reexpression of variables and from the use of special procedures for handling outliers.

Beginning with the milestone "The Future of Data Analysis" that Joe Hodges used in 1962 to kick off his editorship of *The Annals of Mathematical Statistics* (reprinted in Vol. III, pages 391–484), John made a deliberate decision to push, and then push some more, on data analysis. In effect he hitched his star to EDA. John's version of EDA emphasizes pencil and paper plots like stem-and- leaf histograms and box-plots. In this he shows inventiveness of a high order, but with the limitation of primarily addressing small data sets that can be captured by simple graphical summaries carrying direct visual impacts. His work can be seen as a precursor of the large developments called "data mining" and "data visualization," now aimed more at elucidating large and complex data sets, and strongly based in many fields of engineering and physical sciences. These movements aim to create techniques that represent and capture complex highly structured systems. Summarizing and displaying relevant aspects of such data, such as spatio-temporal aspects, is part of analyzing complex systems. Many of John's simple tools will survive especially through widely used plots provided in statistical software packages. There is, however, little attention paid in his writing to the increasing importance of massive computing power in exploring data, and he has little to say about the role of big models that appear to be necessary if complex phenomena are to be understood as a whole, not just piecemeal in small reductive ways.

Although John worked on many important empirical studies, the philosophical volumes bear few traces of his applied work. It is perhaps a consequence of a mainly "bottom up" approach that the larger role of statistics as a driver of broad classes of important issues in science is not assayed. As he himself might have said, we need both "bottom up" and "top down" approaches. Unlike his sometime collaborator Fred Mosteller, John appears to have done little commenting in print about the part played by data analysis in the findings of major projects that he advised on. My sense is that he unwisely assumes that it is obvious when a statistical analysis is successful or useful. Experience in life, perhaps especially in statistical life, teaches that there is rarely agreement about such matters. A real case study is not complete without a presentation of alternatives considered and tried in the course of design and analysis, of challenges to findings following publication, and of lessons learned under criticism. Perhaps it is time for an evaluation of how the EDA movement has led in more and less productive directions, as seen from inside and outside the movement.

John addresses the balance between "exploratory" and "confirmatory" in a short 1980 note in *The American Statistician*,

> We need both exploratory and confirmatory . . .

and again in JWT86 (page xl) where he defends the tilt in his writing toward the former as appropriate for the profession at the time of writing, while worrying

that readers in coming years might read an unintended prejudice. This shows John as fundamentally conservative, aiming to cast himself as contributing to the ongoing development of the discipline, rather than viewing EDA as a new wave that would replace a current paradigm. Here again I perceive the lack of a "top down" explanation. Just what did he think was the message for the practitioner in a reported formal inference such as a p-value? What is "confirmed," and how? I will return below to vexing questions surrounding Bayes and subjective probability, and Fisher's "fiducial" probabilities in relation to Neyman's confidence procedures.

3. Conclusions and decisions. In JWT86 John makes a point of advising the reader to:

> . . . please read chapter [6], "Conclusions vs. decisions" so that we can all have in mind a clear distinction between conclusions accepted on a continuing basis, to be used to guide thought or action wherever and whenever applicable [until replaced or corrected], and decisions "to act as if ABC were true."

The message here is that established scientific findings should not be confused with decisions under uncertainty that must often be made in the absence of firm knowledge. Correspondingly the logic underlying the former must not be confused with the logic underlying the latter. There is little doubt that John was predominantly interested in science more than decision-making, except of course for the different sort of decisions that are wise choices of statistical procedures.

In his introduction to Chapter 6, Lyle Jones picks out key quotes including:

> I believe that conclusions are even more important to science than decisions,

and

> . . . conclusions must be reached cautiously, firmly, not too soon and not too late . . .

and

> . . . must be of lasting value, but not necessarily of everlasting value . . . . And they must be judged by their long run effects by their "truth" not by specific consequences of specific actions.

Jones has further perceptive remarks on Chapter 6:

> Throughout, Tukey emphasizes the importance of separating decision elements from conclusion elements in statistical methods. A test of significant difference is viewed as a qualitative conclusion procedure, while interval estimation is a quantitative conclusion procedure. Tests of hypothesis may be either decision procedures or conclusion procedures; the entwining of decision elements and conclusion elements is a source of confusion to be avoided as much as possible.

Writing in 1955, John is commenting on the relatively new dominance in mathematical statistics of the frequentist decision theoretic framework due to Abraham Wald. In an interesting Appendix 3 to Chapter 6,

> What of tests of hypotheses?

he draws attention to the inappropriateness of Neyman's theory with its

> 5%, !%, and the like

for operational decisions, and basically endorses Wald's theory as encouraging

> a much wider variety of specifications

and showing

> that one should expect mathematics to provide, not a single best procedure, but rather an assortment of good procedures (e.g., a complete class of admissible procedures) from which judgment and insight into a particular instance (perhaps expressed in the form of an a priori distribution) must be used to select the "best" procedure.

In the 1961 "Badmandments" paper that occupies fully 200 pages (Vol. III, pages 187–390), John addresses a confusion that is endemic among otherwise able scientists between a p-value and a posterior probability. As "badmandment #100" John points to the

> classical fallacy of significance testing,

namely,

> The significance level tells you the probability that your result is WRONG.

It remains true today that

> every statistician spots some form of this badmandment quite frequently, and marks up its appearance as an error.

I believe that Jerzy Neyman and Egon Pearson deserve some of the blame for this situation through casting their theory of hypothesis testing as deciding between null and alternative hypotheses, whence it is a short step to misinterpreting a quoted tail area as adjudicating between null and alternative. Fisher was on safer ground in insisting that p-values have nothing to do with decisions, indeed do not involve any formulation of an alternative in their computation, but rather are only indicators of the occurrence of improbable results computed under the null hypothesis. Of course, one can also say that the scientists who repeat this fallacy bear responsibility for not having better training in the meaning of probabilistic arguments.

Carrying the argument one step further, John goes on to suggest that

> investigators who do use such phrases do seem to interpret data in about the same way as those who do not.

Does it matter that confidence coefficients or significance levels are routinely misinterpreted, if the wheels of science appear to be greased by these misinterpretations? Sometimes yes, and sometimes no. If the value is 0.001, and is buried in a purely scientific paper that will either stand the test of replication by others or not, there will be little harm. But if the value is 0.2, and is used as a basis for

an operational decision, then the costs arising from misuse may be real indeed, and posterior probability calculations are advisable.

4. Bayes. Some readers may be surprised by the conciliatory approach to Bayesian inference exhibited in JWT86, stressing that the essence of Bayes is in

> bringing in information from other sources,

while criticizing the

> hope (or even a claim) that Bayesian approaches can also serve us well when we do not wish to draw on any source of information other than the data before us,

as by the use of

> "gentle" or "uninformative" priors.

Two pages later he avers that "personal probability" could be useful for "discriminating judgments" by trained individuals making use of the "probability calculus." He stated that he himself did not work in areas that seemed to need priors or personal probabilities.

He faults the Bayesian movement in statistics as offering

> no really effective place for robustness

and argues that

> Bayesian analysis, as discussed today, tempts us to:
>
> - discard some plausibly important insights and experience, since it is hard to distill them into quantitative form;
> - avoid thinking about systematic errors, since it is not usual to include them in Bayesian models;
> - neglect whatever experience or insight that no one sees how to distill.

These remarks come from the mid-1980s, just as practical Monte Carlo methods for sampling nontrivial posterior distributions were poised to come on stream. Much experience gained since then might have partially allayed John's concerns. In particular, since multiple versions of a Bayesian analysis are easily carried out, Bayesian robustness is routinely assessed through sensitivity analyses that focus naturally on meaningful consequences of alternative assumptions. Far from encouraging avoidance of quantifying important factors such as systematic errors, Bayesian methods provide a straightforward way to introduce evidence from sources not reflected in statistical data. Bayesian thinking and analysis should be seen as complementing and interacting with John's preferred modes.

5. Fisher versus Neyman. Early on, John recognized the fundamental importance of Fisher in the growth of the discipline of statistics, seeing in Fisher a model for using high mathematical talent in the service of science. John edited a republication of Fisher's basic statistical papers [Tukey (1950)],

and he encouraged his students at the time to take a balanced view of various differing British and American viewpoints. He was intrigued in particular by Fisher's "fiducial argument" and made a sustained effort over several years in the 1950s to clarify and elaborate his understanding of the concept. The 14 pages of correspondence between Fisher and Tukey reproduced in Bennett (1990) shows them mostly talking past one another, and culminated in Fisher unceremoniously ejecting John from a visit to his home in the summer of 1955, an event that I had no hint of despite my being a graduate student of John's at the time.

Volume VI of John's *Collected Works* contains two long unpublished documents on fiducial methodology. A long manuscript (pages 55–118) dated August 1957 contains the statement:

> We have asserted as a cardinal requirement that statements of probability, including those of fiducial probability, should be verifiable in terms of frequency.

Fisher of course rejected this "requirement" as part of his larger rejection of the whole decision-theoretic apparatus that was the foundation of Neymanian statistics. History shows that mainstream statistics has remained in the grip of the injunction to be "verifiable in terms of frequency," to the point that Fisher's statistical philosophy is largely unknown to recent generations. The injunction has also served as a defense against Bayesian inroads.

Some comments may be in order, because two aspects of frequency, both important, risk being confused. First, all statistical scientists surely accept that probabilities based in frequencies are the firmest kind. But in making such a statement we are accepting that the two concepts are different, for otherwise the statement is vacuous. So the importance of frequency as a source leaves open the question of what probability "is." The second kind of frequentism is the kind that Neyman enshrined in his doctrine of "inductive behavior," now also rarely taught and largely forgotten, but based on the very much alive principle that choices of statistical procedures should be made by evaluating their performance in the long run. For me, as I assume it was for Fisher, the problem with elevating frequentist verifiability to a *sine qua non* is that there is rarely out there in the objective real world a large set of cases on which verifiability matters, as I think should be increasingly apparent in an era of rapidly increasing complexity of the phenomena we face as a profession. I return below to the practical necessity of making inferences relevant to the specific situation under analysis. All that said, it remains of course theoretically important, and suggestive of consequences of practical choices, to carry out frequentist evaluation of procedures, despite Fisher's intemperate objections to the contrary.

Both in the document cited above, and in the three handouts to the IMS Wald lectures given at MIT in August 1958 (also reprinted in Vol. VI) John attempted to create his own version of the mathematization of the fiducial argument. One technical issue concerns the nonuniqueness of pivotal quantities, including different choices based on minimal sufficient statistics of the same dimension

as the parameter space. Tukey's frustrations with Fisher no doubt reinforced the tendency emanating from numerous other mathematical statisticians whose work Fisher had peremptorily dismissed that Fisher was a crank who would not admit errors in his work. I did not take away this attitude from John's work, however, and went on to take Fisher as a major inspiration for my own work.

The Wald lectures show John as a believer that mathematical analysis outside the decision-theoretic framework might rescue Fisher's fiducial idea from its dominance by Neyman's frequentist reformulation into the confidence argument. Some of Tukey's ideas here, using group invariances as a foundation for some of Fisher's examples, resurfaced in Donald Fraser's "structural inference" theory, but as far as I know, John never took Fisher's theory seriously after 1958. My own attempt to build on Fisher's fiducial argument, now widely known as Dempster–Shafer theory in varied literatures largely outside of statistics, did not draw on the Tukey–Fraser approach, but rests on a more general and also mathematically precise integration of propositional logic with degree-of-belief probability.

In a long joint paper with Mosteller ("Data analysis, including statistics," Vol. IV, pages 601–720) that presages their joint 1977 book, the authors address their concern that, as propounded by Neyman and in a host of textbooks, a confidence interval is only a statement about a hypothetical long run. Why then should a scientist accept a confidence statement as saying anything about a specific real world situation under study? Fred and John opine that:

> . . . when there is no extra information about the parameter, and when typicality and absence of selection do not seem to need challenge, there seems to be nothing wrong with attaching a degree of belief to the confidence interval; in fact it is hard to see why there should be a serious concern.

The side condition "typicality and absence of selection" is very close to Fisher's "no recognizable subset" condition for fiducial validity [Fisher (1956)]. If a confidence statement says nothing specific about a situation at hand without an important added qualification, why is this point not made clear to users inside and outside statistics proper? There is an unresolved problem here that remains too little addressed.

6. In conclusion. Two quotes may serve to back up a contention that John was primarily interested in getting the science right, rather than just contributing to specialized branches of statistics. One piece of advice is: "Unsolved problems of experimental statistics" [(1954) pages 77–105 of Vol. III],

> Statisticians must face up to the existence and varying importance of systematic error

the point being that what is not in data can be as important as what the statistician can see.

In "Data analysis and behavioral science" [(1961), pages 187–389, Vol. III] we find

> ... the establishment of a causal relation always requires two elements, one empirical, the other theoretical.

This serves as a reminder that what statisticians call causation is qualitatively different from scientific causation. The former is really about prediction, such as making predictions about what effects different treatments will have on defined populations and subpopulations. Statistical causation raises important concerns about the advisability of randomization and the dangers of observational studies, issues that can be obscured by the emphasis in many areas of applied statistics on detecting "significant effects." Scientific causation goes a step further, however, as John remarks, and requires the scientist to explain "theoretically" why an effect can be expected, for example, by making a cause plausible through linking to an explanation in terms of a mechanism.

I could add that EDA itself is closer in spirit to how nonstatistical scientists think than to any probabilistic formulations of statistical methods: aiming to provide tools that speak directly to enriching empirical knowledge through computation with raw data.

John's retreat to a primary emphasis on EDA represents a pessimistic view that statistical theory is far from achieving respectability in large areas of science. He saw statistical modeling and uncertain inference as failing to meet standards. Regarding inference in particular, he remarked in the foreword to Volume VI (1990):

> Today I do not believe there is any chance for a single unifying approach to inference.

My own view is more optimistic. While we need to pay close heed to the important issues raised by John Tukey over many years, progress toward greater consensus is possible, requiring mainly that new generations of Ph.D. statisticians adopt more catholic attitudes freer from dated ideologies of decades long past. This is a topic for another time and place [Dempster (1998, 2002)].

REFERENCES

BENNETT, J. H., ed. (1990). *Statistical Inference and Analysis*: *Selected Correspondence of R. A. Fisher*. Oxford Univ. Press.

DEMPSTER, A. P. (1998). Logicist statistics I. Models and modeling. *Statist. Sci.* **13** 248–276.

DEMPSTER, A. P. (2002). Logicist statistics II. Inference. Submitted.

FISHER, R. A. (1956). *Statistical Methods and Scientific Inference*. Hafner, New York.

MOSTELLER, F. and TUKEY, J. W. (1968). Data analysis, including statistics. In *Handbook of Social Psychology*, 2nd ed. (G. Lindzey and E. Aronson, eds.) **2** 80–203. Addison-Wesley, Reading, MA.

MOSTELLER, F. and TUKEY, J. W. (1977). *Data Analysis and Regression*: *A Second Course in Statistics*. Addison-Wesley, Reading, MA.

TUKEY, J. W., ed. (1950). *R. A. Fisher's Contributions to Mathematical Statistics*. Wiley, New York.

TUKEY, J. W. (1954). Unsolved problems of experimental statistics. *J. Amer. Statist. Assoc.* **49** 706–731.
TUKEY, J. W. (1962). The future of data analysis. *Ann. Math. Statist.* **33** 1–67.
TUKEY, J. W. (1980). We need both explanatory and confirmatory. *Amer. Statist.* **34** 23–25.
TUKEY, J. W. (1986). *The Collected Works of John W. Tukey III. Philosophy and Principles of Data Analysis*: *1949–1964*. Wadsworth, Belmont, CA.
TUKEY, J. W. (1986). *The Collected Works of John W. Tukey IV. Philosophy and Principles of Data Analysis*: *1965–1986*. Wadsworth, Belmont, CA.
TUKEY, J. W. (1990). *The Collected Works of John W. Tukey VI. More Mathematical*: *1938–1984*. Wadsworth, Belmont, CA.

DEPARTMENT OF STATISTICS
HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS 02138
E-MAIL: dempster@stat.harvard.edu