The Annals of Statistics
2002, Vol. 30, No. 6, 1640–1648

JOHN W. TUKEY'S CONTRIBUTIONS TO ROBUST STATISTICS

BY PETER J. HUBER

We summarize John W. Tukey's contributions to robust statistics, separating them into four categories: conceptual; tools; techniques; procedures.

In robustness, as in every area he touched, John Tukey produced hundreds of original ideas, some brilliant, fundamental and lasting, some ephemeral. He presented them in a rambling fashion, in survey articles, in technical reports, in mimeographed draft memoranda and some only in lectures. Along the route, like The Three Princes of Serendip (a favorite fable of his), he would make unexpected discoveries by accident and sagacity. When I started working in robustness in 1961, I much profited from the superb reprint and preprint collection then housed in the coffee room of the Berkeley Statistics Department, containing a fair number of his unpublished papers. Tukey loved words, especially those he had created himself, and his baffling terminology made those papers hard to understand, especially to an outsider like me. The precise meaning of his words sometimes changed over time, and important ideas occasionally got lost in the passage from a preliminary to the final version of a paper.

In Tukey's *oeuvre*, his contributions to robustness are among the least organized, and, regrettably, there is still no corresponding Collected Works volume. I shall try to identify some of Tukey's more important or fertile contributions and to separate them into four categories: *conceptual*; *tools*; *techniques*; and *procedures*. The separation between the latter three categories admittedly is somewhat fuzzy, but it clarifies matters if we do not throw things such as diagnostic tools, Monte Carlo techniques and estimation procedures all into the same pot. At the same time I shall try to transmit some impression of John's idiosyncrasies and style of interaction. The choice and assessment of the importance of his contributions are mine and sometimes would differ from his own.

Conceptual contributions. In robustness, Tukey's most decisive contribution was his clear conceptual recognition of the main underlying problem. He apparently was the first to recognize the extreme sensitivity of some conventional statistical procedures to seemingly minor deviations from the assumptions, and he elaborated on this theme in his 1960 paper, "A survey of sampling from contaminated distributions." He was not the first to use scale mixtures of normals to model the distribution of measurement errors [cf. Newcomb (1886)], nor was he

Received September 2001; revised March 2002.
AMS 2000 subject classifications. 62F35, 62F40, 01A70, 62–07.
Key words and phrases. Robustness, resistance, data analysis, Jackknife, sensitivity curve, median polish.

the first to notice that in practice the mean absolute error might be superior to the "optimal" mean square error [cf. Eddington (1914)], to mention just two authors. But he put the finger on the problem, namely on the excessive sensitivity of some classical procedures to seemingly negligible deviations from the distributional assumptions, and he gave an eye-opening quantitative example. Implicitly, his paper made clear that robustness was a stability or continuity problem, but with topologies somewhat different from those hitherto considered in statistics, and that contamination models provided very handy formalizations. To quote some of his own words: "The problem is a large sample problem." "Moreover, since gross non-normality will be detected, practical concern must be focussed on the effects of indetectable or barely detectable non-normality. We shall learn that such effects may be large" [Tukey (1960), page 453]. It is curious but typical that Tukey, despite his background in point-set topology (remember that a version of the axiom of choice has been named after him) left it to others to draw and explicate the not exactly straightforward topology connection.

While Hampel and myself did so in the 1960s, Tukey's own robustness interests shifted to new grounds, namely to heavy-tailed distributions and small samples. In the Princeton robustness study of 1970–1971 [see Andrews et al. (1972)] his principal aim was to find compromise estimators that would behave well over the entire range between the normal and very heavy-tailed distributions. I remember the amount of persuasion it took me to have him consent to the inclusion of T3 (the t-distribution with three degrees of freedom) in that study. I argued that this was advisable in order to populate a hole left in the design; he felt that adding T3 would waste a slot, it being too close to the normal. At a statistics meeting a few months later, a senior statistician then told me that in his opinion T3 was too long-tailed to be of practical interest! I still believe the truth is somewhere in between. Tukey very often, deliberately or not, overstated his points for the sake of the argument. His (over)emphasis on heavy tails, in combination with the paper by Donoho and Huber (1983), later contributed to the exaggerated importance some people assigned to high breakdown point methods in regression. I am sure that Tukey would share my opinion that uncritical insistence on a high (close to 50%) breakdown point is just as foolish as uncritical use of a low (below 5%) breakdown point procedure.

Generally, and especially after he had decided that his central interest was in data analysis [cf. Tukey (1962), page 2], he would rely on his intuition and profess open contempt of mathematics. He would cow weaker discussion partners, including many of his graduate students, by the sheer weight of his authority (and unfortunately impress upon them that mathematics was unnecessary). Nevertheless, under duress he would produce nontrivial mathematical arguments. For example, during the academic year 1970–1971 in Princeton he obviously relished the presence of hard sparring partners (Bickel, Hampel and myself) and would bounce his new ideas off us. When we did not buy one of them, he usually would either retreat gracefully or come forward with a piece of sophisticated mathematical reasoning. We never knew whether the latter had been sitting there all along in the subconscious back of his mind, or whether he had fashioned it on the spot.

Just before Christmas 1970 David Andrews proposed a collaborative undertaking in order to foster interaction between the robustniks then present in Princeton. We all knew that only a study narrowly limited to an area we believed to understand, but which still presented open finite sample problems, had a chance of succeeding in the short time available (till early June 1971). I was then working on the asymptotic theory of M-estimates of regression, and I think that all of us in principle were more interested in robust regression and analysis of variance than in the simple one-dimensional location problems targeted by Andrews' proposed study [cf. Hampel (1997) for further details]. Frank Hampel just has reminded me of a talk by John Tukey on robust regression, commenting on his depth and breadth of insight, and regretting that nobody seemed to have followed up on the full range of open problems raised in that talk. In subsequent years Tukey repeatedly tried to organize a successor study on robust regression along similar lines. However, the goals of such a regression study are elusive to nail down. In particular, because of the lack of permutation symmetry, suitable experimental designs are much harder to devise than in the simple location case. In addition, it would have been necessary for the protagonists to be physically together again for an extended period of time, which was not practicable. These efforts faded when Tukey's focus of interest moved more fully into exploratory data analysis.

Tools. The main obstacle to a more widespread use of nonstandard statistical procedures was (and still is) the difficulty of understanding what is going on, and of determining what they actually do in a quantitative sense. It is necessary to be able to do this not only in general (i.e., for an abstract distributional model), but also in each particular case, with actual data. One desperately needs tools to assist one's understanding (and even more desperately, the ability to use the available tools sensibly). This applies in particular to all robust procedures, which by necessity are nonlinear, and which often are defined through obscure algorithms. Specifically, one needs tools to assess the influence of individual data values on a statistic of interest, as well as tools for estimating its variability (i.e., a substitute for the classical "standard error").

In 1958 Tukey had realized that a method for bias reduction proposed by Quenouille (1956) could be used for precisely those purposes. To emphasize that he regarded it as a crude and simple, universally useful tool, he gave it the colorful name Jackknife.

DEFINITION. Let $T_n = T_n(x_1, x_2, \dots, x_n)$ be an arbitrary statistic. Then the ith jackknifed pseudovalue is defined as

$$T^*_{ni} = nT_n - (n-1)T_{n-1}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n). \tag{1}$$

Tukey pointed out that T^*_{ni} as defined in (1) measures the (suitably scaled) contribution of x_i toward T_n, and as such is useful for diagnostic purposes.

Quenouille's original contribution had been to recognize that the arithmetic mean of the pseudovalues

$$T^*_n = n^{-1} \sum_i T^*_{ni}$$

was an estimator of the same quantity as T_n, but had a smaller bias than the latter. Tukey in addition pointed out that

$$\frac{1}{n(n-1)} \sum_i (T^*_{ni} - T^*_n)^2$$

usually is a good variance estimate for both T_n and T^*_n; actually it is slightly better adapted to estimating the variance of the former than that of the latter. The Jackknife is known to fail for statistics that depend only on a few order statistics, like the sample median. In recent years, after computing has become cheap and universally accessible, the Jackknife therefore has lost ground against bootstrap methods.

Around 1970, Tukey proposed a variant of the Jackknife, the *sensitivity curve*, designed to assess the sensitivity of an estimator to the position of observations not present in the sample. This is a finite sample version of Hampel's *influence curve*. Instead of omitting one observation, one adds a virtual observation x to the sample and assesses its influence on the estimate by

$$SC_{n-1}(x) = n\bigl(T_n(x_1, \dots, x_{n-1}, x) - T_{n-1}(x_1, \dots, x_{n-1})\bigr).$$

The sensitivity curve also illustrates two idiosyncrasies of John. First, some of his more important ideas were propagated through folklore rather than through his printed work: the sensitivity curve is one of several important ideas appearing in the mimeographed preliminary version of *Exploratory Data Analysis* [EDA; Turkey (1970); cf. the reference in Andrews et al. (1972), page 96], but not in the final printed version [Turkey (1977)]. Second, it highlights Tukey's emphasis on the nonprobabilistic aspects of statistics and data analysis: he preferred to rely on the actual batch of data at hand rather than on a hypothetical underlying population of which it might be a sample. His 1960 paper on sampling from contaminated distributions still had freely used the notion of "robustness" (insensitivity of the asymptotic *distribution* of the estimate to small changes in the underlying *population*). But his later work almost exclusively favored the notion of "resistance" (insensitivity of the *value* of an estimate to small changes in the underlying *sample*). Fortunately, it follows from a theorem of Hampel that the two concepts for most practical purposes are equivalent [see Huber (1981), pages 7 and 41].

Techniques. It is difficult to compare families of estimators across families of distributions, if, as usually is the case, one has to rely on simulation methods. The intrinsic random variability of the simulations easily exceeds the size of the differences one is interested in. We became painfully aware of this in the course of the Princeton robustness study. Tukey at that time concocted sophisticated tricks applicable to error distributions representable as the distribution of random variables of the form $X = Y/Z$, where Y is normal, stochastically independent of Z. In particular he devised "Monte Carlo swindles": make use of information on Y and Z available to the person doing the simulation, but not to the statistician, who is only shown the values of X. Later, he elaborated on another idea: when one is sampling from strictly positive densities, then in principle each sample can come from any such density, but with different probabilities; see the book *Configural Polysampling* [Morgenthaler and Tukey (1991)].

While he used to change his focus of attention every few years, some ideas had long gestation periods. At the 1965 Berkeley Symposium he had drawn a few of us in typical Tukey fashion into an incisive, long and exhausting, but inconclusive, discussion on the advisability of conditioning in robustness and how one might go about it. At that time I did not understand what he was up to (I suspect he did not either), nor how it would be possible to condition in the robustness context, where one lacked a precise, unique model. Of course his concern about conditioning had much to do with his preference for dealing with actual data batches of moderate size rather than with hypothetical underlying distributions, but I believe now that this discussion already contained the seed for configural polysampling.

Procedures. The separation between exploratory data analysis and robustness is naturally blurred. Tukey proposed many quick and dirty procedures for analyzing data with pencil and paper, all of them robust. Examples are the trimeans (weighted means of three selected order statistics) and several robust straight-line fits and robust smoothers. He himself loved to analyze data by hand. In seminar talks, he usually would sit in the back row, occupying his hands and half of his mind with what he called his "knitting"—a data set he was analyzing by scratching down numbers with a four-color ballpoint pen. The knitting metaphor is most apposite, if one remembers that he particularly loved to do iterated running medians. Interactive use may be more important than the specific procedures themselves: the main purpose of them is to help the data analyst look at the data in many different ways. Tukey sometimes said that his philosophy of data analysis had been expressed already in 1 Corinthians 6.12 ("All things are lawful for me, but not all things are helpful"). For such an approach to be successful, the analyst should share Tukey's own brand of data analytic intuition, which was absolutely uncanny. It was most impressive how he homed in quickly on any peculiarities of a data set.

In Tukey's view, robustness is an attribute of the statistical procedure, typically to be achieved by weighting or trimming the observations. This ought to be contrasted to, say, George Box's view, who thought that the data should not be tampered with and that the model itself should be robust. There was a facetious, but highly illuminating, interchange on this point between the two at an ARO meeting. Box reminded Tukey that he (Box) had invented the term and he could define robustness to be anything he wanted it to be. (I think we have here a question of the chicken and the egg: which is first, a robust procedure or a least favorable model? John Tukey in his later years disliked and avoided models—note the absence of models in his EDA [Tukey (1977)]—while George Box, being a Bayesian, was strongly model-based.) In response to a request of the organizer of that meeting (Robert Launer), John then sat down in his room that night and wrote a short descriptive note on his views on robustness, which was inserted verbatim into the proceedings [Tukey (1979), pages 103–106]. This note is highly interesting, because it contrasts what matters to the user of robust procedures and what to the tool-forger. In Tukey's words: The former needs a *reasonably* self-consistent set of procedures that are *reasonably* easy to use; all efficiencies between 90 and 100% are nearly the same for the user. The tool-forger, on the other hand, should pay attention to another $1/2\%$ of efficiency. Tukey comments that just which robust/resistant methods you use is *not* important—what is important is that you use *some*. It is perfectly proper to use both classical and robust/resistant methods routinely, and only worry when they differ enough to matter. *But* when they differ, you should think *hard*.

Tukey's intuition regarding robust procedures was less sharp than his data analytic intuition. This had less to do with his intuition per se than with the fact that he neglected to keep it in check by (heuristical) mathematical arguments. This also may have been the reason why over the years he proposed and experimented with literally hundreds of procedures. It is little known that the Monte Carlo experiment performed in 1970–1971 and described by Andrews et al. (1972), which had investigated the performances of some 68 location estimators under some 40 different "situations" (i.e., error distributions and sample sizes), was but the A-wave of a sequence of successor experiments. Among them, John had reserved the H-wave to Bickel, Hampel and myself, and we used it to fill two conspicuous omissions of the original study (rank estimates, and outlier rejection rules followed by the sample mean). The H-wave was run in 1972 when Tukey and his co-workers already had progressed several waves beyond, to N or so. I have it from hearsay that he ran out of letters of the alphabet before running out of steam.

In distinction to his contributions to spectrum analysis, most of which are here to stay in their original form, his robust procedures, while providing food for thought, leave room for improvements. A problem with several of them is that he did not bother about details and recommended them for general use prematurely. For example his "Biweight," a still very popular redescending robust estimate of location proposed by him shortly after the Princeton robustness study, does not quite measure up to the redescenders previously designed by Hampel in the course of that study, because it is redescending too steeply in the flanks. Contrary to common belief, the “Biweight” is more arbitrary than Hampel’s estimates, and I guess it became so popular only because it looks nicer (i.e., smoother and simpler). While the flaw might be negligible to the user, it is large enough to irk the tool-forger. In any case, many of his procedures are highly intriguing and challenging. For example, the “Shorth”—the mean of the shortest half of the sample—is intriguing because it has a slower than usual convergence rate (namely $n^{-1/3}$ rather than $n^{-1/2}$; the slower rate had first jumped into our eyes when we looked at the computer outputs). Years later a modified and generalized version was reincarnated by Hampel (1975) and propagated by Rousseeuw (1984) in the form of least median of squares (LMS) regression estimates.

Another problem is that Tukey had defined many procedures through ad hoc algorithms without clearly enunciated goals. Indeed, it appears that he often was more interested in the algorithmic process than in what it ultimately achieved. For example, consider his *median polish*, an appealingly simple iterative method for robustly decomposing a two-way table into (overall) + (row effects) + (column effects) + (residuals):

$$x_{ij} = m + a_i + b_j + r_{ij},$$

such that the row-wise and column-wise medians of the residuals are all 0; see Tukey [(1977), Chapter 11]. In most cases, his procedure stops after a small finite number of steps, and it has the (fairly obvious) property that each iteration decreases the sum of the absolute residuals. It is, however, much less obvious that the process hardly ever converges to the true minimum. As a rule, it stops just a few percent above, but an example (due to Anscombe) shows that the relative deficiency can be arbitrarily large. I guess that most people, once they become aware of these facts, for machine calculation will prefer the L_1-estimate of the row and column effects, which gives a fixed-point of the median polish algorithm achieving this minimum. See in particular Kemperman (1984) and Chen and Farnsworth (1990).

But even when his intuition did not hit the intended nail, it often hit a neighboring one. The interplay between trimming and Winsorizing is a good example. A batch of univariate data is trimmed by dropping a fixed number of extreme values, say g on the left and g on the right. Tukey felt it would be preferable not to drop those values entirely, but merely to reduce their influence. He therefore proposed that, instead of dropping, one should replace those values by the next adjacent order statistic, that is, by the $(g + 1)$st and the $(n - g)$th respectively. In the honor of Charlie Winsor he called this procedure Winsorizing. Interestingly, and rather counter-intuitively, it turned out a few years later that trimming *does* exactly what Winsorizing was *supposed* to do but, on the other hand, that the standard error calculated from the Winsorized sample asymptotically gives the correct value for the standard error of the trimmed mean. See Huber [(1981), pages 58–59].

Common themes. Tukey's contributions to robustness must be viewed in the context of the philosophical issues pervading his work in data analysis. They have been summarized by him in the opening and closing sections of his long paper "The future of data analysis" [Tukey (1962)]. In distinction to some of his later pronouncements, his views there are expressed in a balanced fashion. These sections still are eminently readable and should be required reading for any aspiring statistician; the issues are likely to stay around for a while to come. He pointed out that inference in the sample-to-population sense is only part, not the whole, of statistics and data analysis. Thereby he pushed probability theory away from the center stage it had occupied in statistics for the preceding half of a century. He warned against the dangers of optimization and moved the role of mathematical proof into the background. Instead he stressed the importance of judgment. On such a basis we can for example justify the use of a procedure, robust or otherwise, even if we know its properties only approximately and in particular cannot quantify its accuracy in probabilistic terms (whether these are confidence intervals or posterior probabilities). Also, we are allowed to rely on simulation ("if [it] be used wisely") rather than on rigorous proof. In short, a data analyst should behave like a scientist rather than as a pure mathematician. Of course I should add that, for all this to work, you had better be endowed with a good share of John Tukey's statistical intuition and judgment!

Acknowledgments. I thank David Brillinger, Frank Hampel, Robert Launer and a referee for helpful interchanges and suggestions.

REFERENCES

ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P. J., ROGERS, W. H. and TUKEY, J. W. (1972). *Robust Estimates of Location*: *Survey and Advances*. Princeton Univ. Press.

CHEN, S. and FARNSWORTH, D. (1990). Median polish and a modified procedure. *Statist. Probab. Lett.* **9** 51–57.

DONOHO, D. L. and HUBER, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann* (P. J. Bickel, K. Doksum and J. L. Hodges, Jr., eds.) 157–184. Wadsworth, Belmont, CA.

EDDINGTON, A. S. (1914). *Stellar Movements and the Structure of the Universe.* Macmillan, London.

HAMPEL, F. R. (1975). Beyond locations parameters: Robust concepts and methods (with discussion). *Bull. Inst. Internat. Statist.* **46** 375–391.

HAMPEL, F. R. (1997). Some additional notes on the "Princeton Robustness Year." In *The Practice of Data Analysis*: *Essays in Honor of John W. Tukey* (D. R. Brillinger, L. T. Fernholz and S. Morgenthaler, eds.) 133–151. Princeton Univ. Press.

HUBER, P. J. (1981). *Robust Statistics.* Wiley, New York.

KEMPERMAN, J. H. B. (1984). Least absolute value and median polish. In *Inequalities in Statistics and Probability* (Y. L. Tong, ed.) 84–103. IMS, Hayward, CA.

MORGENTHALER, S. and TUKEY, J. W., eds. (1991). *Configural Polysampling. A Route to Practical Robustness.* Wiley, New York.

NEWCOMB, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *Amer. J. Math.* **8** 343–366.

QUENOUILLE, M. H. (1956). Notes on bias in estimation. *Biometrika* **43** 353–360.

ROUSSEEUW, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79** 871–880.

TUKEY, J. W. (1958). Bias and confidence in not-quite large samples (abstract). *Ann. Math. Statist.* **29** 614.

TUKEY, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics*: *Essays in Honor of Harold Hotelling* (I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow and H. B. Mann, eds.) 448–485. Stanford Univ. Press.

TUKEY, J. W. (1962). The future of data analysis. *Ann. Math. Statist.* **33** 1–67.

TUKEY, J. W. (1970). *Exploratory Data Analysis.* Mimeograph. [Preliminary edition of Tukey (1977).]

TUKEY, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading, MA.

TUKEY, J. W. (1979). Robust techniques for the user. In *Robustness in Statistics* (R. L. Launer and G. N. Wilkinson, eds.) 103–106. Academic Press, New York.

POSTFACH 198
CH-7250 KLOSTERS
SWITZERLAND
E-MAIL: peterj.huber@bluewin.ch