Statistical Science
2003, Vol. 18, No. 3, 319–331

John Tukey and Robustness

Karen Kafadar

Abstract. John Tukey's impact on statistics, and on science in general, is broad and lasting. This article discusses some of these contributions, with a special emphasis on those that led to the development of robust methods and data exploration. In view of today's emphasis on data mining techniques, the recollection of Tukey's influence is especially timely.

Key words and phrases: Robust methods, exploratory data analysis, statistical graphics, biweight, boxplot, Gaussian distributions, long-tailed distributions, contaminated normal distributions, time series.

1. INTRODUCTION

Much of the work in the field that came to be known as "robust methods" was inspired in one way or another by John Tukey. Although the first appearance of the word "robustness" seems to have been in an article by Box (1953), it is clear from Tukey's early work in nonparametric methods and rank-based inference techniques that he had viewed for some time the Gaussian assumption with skepticism. His concerns about parametric inference can be detected in many of his early publications, and explicitly so in his later ones. In this article, I review some of this early work, discuss in particular one inspiring article that was published in 1960 (Tukey, 1960a), describe some of the ways in which he contributed to the field, either directly (through his own publications) or indirectly (through ideas that he sprinkled throughout his articles or in conversations with others) and conclude with some observations about the present impact of robust methods on statistical practice today.

2. GENERAL PHILOSOPHY

Tukey's general philosophy on the need for robustness in data analysis appears in many of his articles, most notably the landmark 1962 article, "The future of data analysis":

> *We need to tackle old problems in more realistic frameworks*. The study of data analysis in the face of fluctuations whose distribution is rather reasonable, but unlikely to be normal, provides many important instances of this. So-called nonparametric methods, valuable though they are as first steps towards more realistic frameworks, are neither typical nor ideal examples of where to stop. . . .
>
> The development of a more effective procedure for determining properties of samples from non-normal distributions by experimental sampling is likely, if the procedure be used wisely and widely, to contribute much to the practice of data analysis. [*Tukey*, 1962, *Section* 1.]

(Tukey often made a point of distinguishing between *experimental sampling*, which he viewed as straightforward simulation, and *Monte Carlo*, which he viewed as a "smart" form of simulation.) In another article much later, he wrote:

> Probability modelers seem to want to believe that their models are entirely correct. . . .
>
> Data analysts regard their models as a basis from which to measure deviation, as a convenient benchmark in the wilderness, expecting little truth and relying on less.
>
> The practitioner/theorist of statistical inference was once supposed to think like the probability modeler, but the rise of robust/resistant techniques and theory presages the day when both practitioners and theorists will speak and act as if the truth were, hopefully, somewhere "not too

Karen Kafadar is Professor, Department of Mathematics, University of Colorado, Denver, Colorado 80217-3364 (e-mail: kk@math.cudenver.edu).

> far" from their models. [*Tukey*, 1979, *pages* 121–122.]

Despite his concerns about parametric inference, Tukey often had some sort of "traditional" method or model in mind, for purposes of comparison. For example, the term "Gaussian deficiency" referred to the loss in efficiency of a procedure, *relative* to the most efficient procedure that one could use for the Gaussian. In trying to understand the motivation for some of Tukey's proposed methods, often it was helpful to recall the method that would be optimal for some standard or traditional model, from which one could often see more clearly the method's inadequacies in the face of departures from that model, and why Tukey's proposal often addressed those inadequacies.

Tukey also often left details to the readers, particularly in later years. Likely, he had already processed in his own mind, at least to some degree, the necessary mathematical derivations, but he left it to others to work out the formal theorems and proofs. [A good example is the biweight; recent work by David W. Scott (2001) shows that the influence function for his L_2E estimator resembles closely that of the biweight, providing some justification for the biweight's remarkable performance in so many contexts, and confirming Tukey's early statement that "the distribution of our observations will often be one for which minimizing a quadratic is a very bad choice" (Tukey, 1965, page 669).]

3. EARLY WORK

Tukey's first publications in statistics (after his mathematical publications in topology) involved nonparametric methods and order statistics. In "A formula for sample sizes for population tolerance limits" (Scheffé and Tukey, 1944) and "Non-parametric estimation. I. Validation of order statistics" (Scheffé and Tukey, 1945), Scheffé and Tukey use the probability integral transform of order statistics (uniform) to derive population tolerance limits and confidence bands for a continuous cumulative distribution function (c.d.f.). The following year, a very short article appeared, "An inequality for deviations from medians" (Tukey, 1946), in which he derived an upper bound for the sum of the absolute deviations from the median: $h(n) \cdot E(\sum |x_i|) \le E(|\sum x_i|)$, where x_i are independent random variables with median 0, and $h(2k+1) = h(2k+2) = (2k+1)!/(2^k k!)^2$. This result suggests that he already recognized the value of least absolute deviations as an alternative to least squares.

In "Low moments for small samples: a comparative study of order statistics," Hastings, Mosteller, Tukey and Winsor (1947) derived the means, variances and covariances of order statistics for small ($n \le 10$) samples from the Gaussian, uniform and "special" distributions (the last later known as a member of the Tukey λ family with $\lambda = 0.1$) and showed that the asymptotic formulas were not trustworthy (except for the standard deviation of central order statistics).

Tukey also realized as early as 1948 the practical difficulties of inference in small samples when distributions were not Gaussian. He opened his article "Some elementary problems of importance to small sample practice" as follows:

> This memorandum outlines a few problems concerning small samples whose solution would have considerable practical application and which seem approachable by some combination of experimental sampling, empirical approximation, and, possibly, analytical investigation.
>
> The first group of these problems is connected with the behavior of t for nonnormal distributions and the possibilities of finding an improved expression for use in such cases.
>
> The second group is related to the "studentization," in a generalized sense, of the sample variance. These problems are generalized to the higher k-statistics.
>
> Finally, the last group of problems is concerned with t and its modifications when the population is finite. [*Tukey*, 1948a.]

This article confirms Tukey's worries about deviations from normality, even before the concept of "robustness" had been mentioned, much less developed. This article also shows his proposal to use three methods (experimental sampling, empirical approximation, analytical investigation) for verification and demonstration, methods on which he relied throughout his career. In the second section of this paper, Tukey recalled the problems of skewness in the distribution of t for underlying skewed populations and proposed a potentially less sensitive "t-statistic":

> It has been appreciated for a long time that two-sided t-tests are insensitive to deviations from normality [robustness of validity]. ...It would be highly desirable to have a modified version of the t-test with

a greater resistance to skewness when used as a one-sided test. Since there is always a *quid pro quo* in statistics (unless inefficient methods are involved), we must expect to lose a little power when the underlying distribution is symmetrical, but this loss may be expected to be very small. A similar allowance for the kurtosis of the sample might possibly be worthwhile if very simple.

Casting about for a possible way in which to bring about this modification, our attention is struck by the work of Egon Pearson (5, 1929, pp. 280–286) and Walsh (unpublished) on the use of the expression

$$\frac{\text{midrange} - \text{assumed mean}}{\text{range}}$$

in very small samples. The efficiency is very high when the underlying distribution is normal, and the distribution is remarkably little affected by deviations from normality. . . .

With this background, we can now formulate some specific problems.

Problem A1: What modifications of t are (i) simple, (ii) distributed with less dependence on the underlying distribution than t, and (iii) have reasonable power? How are these modifications distributed when the underlying distribution is normal?

Problem A2: What modifications of t are (i) simple, and (ii) distributed similarly for normal and binomial underlying distributions? How are these modifications distributed when the underlying distribution is normal?

Problem A3: Are there general procedures for finding such modifications of given expressions, which are less dependent on the underlying distribution? [*Tukey*, 1948a, *pages* 205–206.]

Along these lines, much later, Arthur (1979) investigated the behavior of Student's t-statistic under skewed and stretched-tailed distributions, and Horn (1983) developed some t-like statistics, similar to Tukey's proposal above but based on only two or four order statistics. [Incidentally, the article immediately preceding Tukey (1948a) was written by Charles Winsor, to whom Tukey often referred as his mentor; it discussed the value of transformation to ensure additivity in an analysis of variance, specifically the probit transformation for data in the form of proportions (Winsor, 1948). Transformation (or re-expression) in general is a common theme in Tukey's work. See Section 7 below and the article by Hoaglin (2003)].

Another useful article that appeared in the same year concerned the issue of misweighting when combining sample means of various precisions for an overall estimate of a common population mean. He presented the results of this very short article, "Approximate weights" (Tukey, 1948b), in his legendary "Statistics 411" course (purportedly for seniors, but attended by graduate students and visiting professors as well). This article provides a very useful bound to answer the question: how "unoptimal" is a misweighted average? Given observations $x_1, \dots, x_n$ with associated variances $\sigma_1^2, \dots, \sigma_n^2$, the optimal weighted average uses weights $w_i^* = 1/\sigma_i^2$, $i = 1, \dots, n$ (the optimal weighted median uses weights $w_i^* = 1/\sigma_i$). But what if one uses instead the wrong weights w_i, $i = 1, \dots, n$? Tukey defined R as the ratio of the maximum to the minimum of the ratios between the contemplated weights and the optimal weights, that is,

$$R = \frac{\max_i (w_i / w_i^*)}{\min_i (w_i / w_i^*)}.$$

Then he derived the following bound for the ratio of the misweighted mean to the optimally weighted mean:

$$\operatorname{Var}(\bar{y}_m) / \operatorname{Var}(\bar{y}_*) \le 1 + (R-1)^2/(4R);$$

that is, small variations in weights generally cause small changes in the variances of the weighted average. In conveying to the students the consequences of this result, Tukey said:

- "$R = 2$, don't worry" (this corresponds to only a 12.5% increase in the variance);
- "$R = 3$, you can live with" (33.3% increase);
- "$R = 4$, you begin to get queasy" (56.25% increase).

Tukey relied on order statistics again in deriving an approximate test for the difference in location between two populations in "A quick, compact two-sample test to Duckworth's specifications" (Tukey, 1959). This test, apparently derived for a client named Duckworth, proceeds by counting the number of "overhanging" observations (i.e., the number of observations in sample 1 that are less than the minimum in sample 2, plus the number of observations in sample 2 that are greater than the maximum in sample 1). If this count is 7, 10 or 13, then this test statistic indicates different population means at the 5%, 1% or 0.1% level of significance, when the sample sizes are approximately equal. (Tukey derives adjustments to the critical levels 7, 10, 13, if the ratio of sample sizes exceeds 1.33.) He justifies this rather simple test on the basis of "practical power": "If [a procedure] is compact, then it can afford somewhat reduced power, for what is practically important may be, roughly, the *practical power* of the procedure in the sense of Churchill Eisenhart, who has defined practical power as the product of the mathematical power by the probability that the procedure will be used" (page 32). About one-half of this article is devoted to the combinatorial derivations that justify the error rates in the procedure. The test is enormously useful and is independent of parametric assumptions.

In perhaps the most widely cited abstract of all time, Tukey described, in only seven sentences, a procedure for assessing the uncertainty in a parameter estimate (Tukey, 1958):

> The linear combination of estimates based on all the data with estimates based on parts thereof seems to have been first treated in print as a means of reducing bias by Jones (*J. Amer. Stat. Assn.*, Vol. 51 (1956), pp. 54–83). Let $y_{(\cdot)}$ be the estimate based on all the data, $y_{(i)}$ that based on all but the ith piece, $\bar{y}_{(i)}$ the average of the $y_{(i)}$. Quenouille (*Biometrika*, Vol. 43 (1956), pp. 353–360) has pointed out some of the advantages of $ny_{(\cdot)} - (n-1)\bar{y}_{(i)}$ as such an estimate of much reduced bias. Actually, the individual expressions $ny_{(\cdot)} - (n-1)\bar{y}_{(i)}$ may, to a good approximation, be treated as though they were n independent estimates. Not only is each nearly unbiased, but their average sum of squares of deviations is nearly $n(n-1)$ times the variance of their mean, etc. In a wide class of situations, they behave rather like projections from a non-linear situation on to a tangent linear situation. They may thus be used in connection with standard confidence procedures to set closely approximate confidence limits on the estimand.

When asked how the name "jackknife" came to be, Tukey responded, "If you had exactly the right tool for the job, you'd use it. But if you don't, then you'd use a jackknife." (See also Mosteller and Tukey, 1968, page 655; Mosteller and Tukey, 1977, page 162.) So, if we knew the distribution of $y_1, \ldots, y_n$, then we could derive, in theory or via simulation, the distribution of T, and hence 95% confidence limits or any other measure of uncertainty. But, since we do not, the jackknife offers an all-purpose, handy and useful tool. Twenty years later, Efron proposed the bootstrap, which has the same flavor as the jackknife but which has different properties (Efron, 1982, 1992; Efron and Tibshirani, 1993). The measure of uncertainty provided by the jackknife is *not* robust to outliers. Mosteller and Tukey (1977, pages 140–141) provide an illustration where all but one of the pseudovalues have nearly the same magnitude, and the one pseudovalue without the outlier is much different. [In fact, the pseudovalues were found to relate directly to the sensitivity curve of an estimator, a finite sample version of the influence function (Andrews et al., 1972); more discussion on this point appears in Hampel et al., 1986, page 95.] The robustness of the jackknife variance led Tukey to collaborate in one of his last projects with Luisa Fernholz and Stephan Morgenthaler (Fernholz, Morgenthaler and Tukey, 2004) on the "multihalver jackknife," a version of the jackknife in which subsamples are selected in accordance with Hadamard matrices (on which Plackett–Burman designs are also based). The balance in the constructed subsamples leads to improved performance in variance estimates and to methods for outlier detection. (Luisa Fernholz saved Tukey's handwritten "flow-chart" that served as the basis for this work and that showed the extent to which the details were already well formulated in his own mind.)

4. TRIMMING AND WINSORIZATION

The first appearance of the word "robustness" may have been in an article by G. E. P. Box entitled "Non-normality and tests on variances": "It would appear, however, that this remarkable property of 'robustness' to non-normality... is not necessarily shared by other statistical tests, and in particular is not shared by the tests for equality of variances" (Box, 1953). This article concludes, "The property of robustness I believe to be even more important in practice than that the test should have maximum power and that the statistics employed should be fully efficient. Where necessary I believe that the latter qualities should be sacrified to ensure the former.*" The asterisk refers to a footnote at the end of the article: "* Since writing the above a very interesting paper by J. W. Tukey (1948) has come to my notice which has many points of contact with the present paper and which expresses similar views to the above." [The cited article was Tukey (1948a) in *Human Biology*, quoted earlier.] In a later publication, "Permutation theory in the derivation of robust criteria and the study of departures from assumptions," Box and Andersen (1955) studied whether departures from the assumed Gaussian distribution affected the size of conventional test statistics, such as that used in Student's t-test or in the F-statistic in the analysis of variance (i.e., was the actual type I error rate still the same as the nominal one?). For long-tailed distributions, they showed that these tests were conservative (i.e., the actual error rates were *less* than the nominal error rates), while the opposite held true for short-tailed distributions. "In practice, long tails seem more frequent than short" (Tukey, 1960a, page 458), so presumably statisticians had no reason to worry about using their traditional methods in non-Gaussian situations, so long as the distributions were symmetric and not short-tailed. Despite the appearance of both the word and the concept in the literature, a formal definition and operating framework for "robustness" was still needed.

A critical leap forward into the development of robust methods started with Tukey's paper, "A survey of sampling from contaminated distributions," that appeared in a volume dedicated to Harold Hotelling (Tukey, 1960a). Prior to its publication, Edgeworth threw the statistical community into a debate when he proposed the mean absolute deviation from the mean as a reasonable estimator of scale. Fisher seemingly halted this debate by publishing the "best asymptotically normal" results of maximum likelihood estimators, thereby demonstrating the superiority of the estimator based on the usual sample variance (David, 2002). Tukey reopened this debate when he investigated the performance of the sample mean and sample standard deviation on contaminated normal distributions of the form $(1-\varepsilon)N(0,1)+\varepsilon N(0,9)$ as the fraction ε of the contaminating normal distribution, having three times the standard deviation of the target normal, varies from 0 (no contamination) to $\frac{1}{2}$. Several graphs show an in-depth analysis of the degradation in performance of traditional estimators for the Gaussian distribution under the contaminated normal model. In the introduction to that paper, Tukey presents his reasons for investigating a trimmed mean rather than the ordinary sample mean as an estimate for the population mean, and a Winsorized variance rather than the sample variance to estimate the population variance:

> Some years of close contact with the late C. P. Winsor had taught the writer to beware of extreme deviates, and, in particular, to beware of using them with high weights. Using second moments to assess variability means giving very high weights to extremely deviant observations. Thus the use of second moments, unquestionably optimum for normal distributions, comes into serious question. When this point was raised in conversation, real differences of opinion between some of the statisticians concerned showed themselves. The earliest work on sampling from contaminated distributions was carried out in an attempt to develop facts which would help to quiet this clash of opinion. [*Tukey*, 1960a, *pages* 449–450.]

Tukey frequently acknowledged his debt to Charles Winsor as one of the people "from whom the author learned much" [dedication of *Exploratory Data Analysis* (Tukey, 1977)]. In an interview, he said, "associating with Charlie and living in the data-rich environment where what we were doing was trying to make sense out of the data left me with an ultimate data orientation" (Fernholz and Morgenthaler, 2000, page 83). Tukey cited Winsor as the namesake of the procedure that has come to be called "Winsorization" in a discussion of an article that appeared also in May 1960:

> Finally, there is Winsorization. Charles Winsor put forward a principle of quite general application, namely: While the numerical value of an apparently wild observation is untrustworthy, the direction of its deviation (e.g. high or low) is worthy of attention. He applied this principle to outliers by taking the largest deviations or largest residuals and decreasing their magnitude, while retaining their sign, until they are equal to the next largest ones, thus making a qualitatively reasonable adjustment. In fact, it is possible to give quite quantitative reasons why this sort of adjustment is not only a convenient approach but an effective one. [*Tukey*, 1960b, *page* 160.]

[Tukey often proposed new ideas in his brief "discussions" of other papers; for other examples, see Tukey (1961, 1979).] W. J. Dixon, who worked with Tukey at Princeton, also wrote on the use of Winsorized estimates of location and scale in univariate samples (Dixon, 1960).

The landmark article for robust methods (Tukey, 1960a) proposed much more than just Winsorization.

Tukey also addressed the robustness issue raised by Box and Andersen (1955) from the perspective of *power*:

> It would seem that questions of robustness of efficiency are intrinsically at least as important as questions of robustness of significance levels. And, as soon as it develops, as it shortly will, that failures of robustness of efficiency may be very substantial, the need for more work on the robustness of efficiency will be clear and pressing. [*Tukey*, 1960a, *page* 450.]

The article also provides, in a style unique to Tukey, the reasons why robust approaches must be considered:

> **THE FIRST QUESTION**
>
> Given two normal populations with the same mean, one having three times the standard deviation of the other, it is proposed to prepare a sequence of mixed populations by adding varying small amounts of the wider normal population to the narrower one. It is well known that, in large samples, the relative efficiency as a measure of scale of the mean deviation compared with the standard deviation is 88% when the underlying population is normal. As specific amounts of the wider normal population are added to the narrower one, thus defining new classes of distributions of fixed shape, will the relative efficiency for scaling of the mean deviation compared to the standard deviation increase or decrease?
> Notes:
>
> 1. The possible answers are "increase," "stay the same," and "decrease."
> 2. The problem is a large-sample problem.
> 3. To find the answer, after giving the question long and careful thought, turn over two pages.

Tukey continued in this vein with "The first answer," "The second question," "The second answer," "The third and fourth questions" and "The third and fourth answers," leading to the conclusion that the mean absolute deviation from the mean will surpass the sample standard deviation as an efficient estimate of the population standard deviation when the contamination is as little as 0.018 [the correct value, due to Huber, is actually 0.0018; cf. Huber (1981), page 3]; that is, with an average of only 2 outliers in a sample of size 1000, where on average 998 come from $N(0, 1)$ and 2 from $N(0, 9)$, already the mean absolute deviation from the mean surpasses the sample standard deviation. This seminal paper highlighted many key concepts that directed the research in robust methods for several decades:

(1) robustness of efficiency versus validity;
(2) evaluation on contaminated normal distributions, $(1-\varepsilon)\cdot N(0,1)+\varepsilon\cdot N(0,9)$, which later led to the consideration of the "one-wild situation," which has exactly $n-1$ observations from $N(0, 1)$ and 1 from $N(0, 100)$;
(3) measures of "deficiency," such as

$$\text{“deviation ratio”} = \frac{\text{(deviation of \%-point from mean)}}{\text{(deviation of \%-point from mean for normal)}};$$

(4) long-versus short-tailed distributions (Tukey, 1960a, page 458);
(5) appropriate scaling of distributions for purposes of comparison;
(6) reliance on asymptotic results, because:
   (i) "Both the asymptotic average value... and the asymptotic variance of [the average] exist for 'reasonable' distributions...
   (ii) [Finite sample averages and variances] are deliberately not defined exactly...
   (iii) Only the distributions of the original [random variables] appear in the definitions...
   (iv) The so-called 'Δ-process,' or 'method of propagation of error,' is usually valid in terms of these asymptotic techniques" (Tukey, 1960a, page 459).

   [Later, Morgenthaler and Tukey (1991) and coauthors developed a theory and an approach for constructing robust procedures that achieved finite-sample robustness using configural polysampling, because "an asymptotic theory can give very useful indications at its best and can be misleading at its worst" (Morgenthaler and Tukey, 1991, page 3).]

As a result, the field moved forward in its consideration of:

- both types of robustness;
- efficiency measures quoted with sufficient accuracy to be trusted, but no more ("there is little if any sense in paying attention to an efficiency figure to greater precision than $\pm 10\%$ of itself" (Tukey, 1960a, page 473);

- other robust estimators of both location and scale, because "if contamination is a real possibility (and when it is not?), neither [sample] mean nor [sample] variance is likely to be a wisely chosen basis for making estimates from a large sample... [due to] nearly imperceptible non-normalities" (Tukey, 1960a, page 474).

Tukey concluded this article by recommending the use of the trimmed mean and mean absolute deviation as a "frequently useful compromise" for estimating location and scale, and that "more work is needed" (Tukey, 1960a, page 485).

Tukey and his coauthors continued to study the use of trimmed means and Winsorized standard deviations in a "t-like" statistic in two further articles (McLaughlin and Tukey, 1963; Dixon and Tukey, 1968). Other articles using trimming and Winsorization for two samples later followed (e.g., Yuen and Dixon, 1973; Yuen, 1974).

5. THE PRINCETON ROBUSTNESS STUDY

The academic year 1970–1971 provided the opportunity to study these issues. During that year, plans were formulated to execute a large-scale, highly efficient Monte Carlo simulation of the performance of 64 estimators of location for the symmetric one-sample location problem. These estimators included M-, L- and R-estimates, many of which were proposed and tested specifically for this simulation (e.g., hampels). For reasons given in Tukey (1960a), only long-tailed departures were considered, and, for reasons of simulation efficiency that dictated the use of a Monte Carlo swindle [which was used also in Dixon and Tukey (1968) and nicely explained in Simon (1976)], only distributions of the form N/I were considered, where N represents a standard Gaussian random variable and I represents a positive random variable that is independent of N. This class includes a variety of long-tailed distributions, depending upon the distribution of I; for example, Normal (I is a constant), Cauchy (I is another standard Gaussian), "Slash" (I is uniform on $[0, 1]$, which creates a random variable whose distribution is, as Tukey often said, "as unrealistic as the Gaussian but in the opposite direction"), Student's t (I is $\sqrt{\chi_k^2/k}$), and the contaminated normals (Tukey, 1960a). [See also Rogers and Tukey (1972) for an in-depth discussion of this class.] Notice that the one-wild situation is not a member of this class, since *exactly* one observation comes from N(0, 100), not a probabilistic fraction $1/n$ (which results in a random number of "outliers" in a given realization). This study also led to the notion of the three "corners" to represent plausible situations that might be encountered in practice [extremely optimistic (Gaussian), extremely unrealistic (Slash) and one outlier (One-wild)]. The outcome was the book *Robust Estimates of Location*: *Survey and Advances*, published in 1972 and co-authored by D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers and J. W. Tukey. (This book is often cited as "Andrews et al., 1972", but David Andrews once cited it to me as "et al. Tukey.")

One of the important consequences of this book was the use of simulation (or, preferably, a more efficient Monte Carlo swindle) as an acceptable form of "proof." Many years would have to pass before simulation would become acceptable as a demonstration of performance; even in 1972, many skeptics still held the belief that only analytical theory could be used to justify performance. Today, thirty years later, that view has changed greatly, and many publications in statistics utilize simulation in some form, either as an intrinsic part of the methodology itself or for purposes of evaluating it. But Tukey's advice to use *efficient* simulation still holds today as much as it did then.

Another important consequence of this book derived from the realization that M-estimates of location can be computed as iteratively reweighted means, presaging the use of iteration in many current statistical procedures used today. Since it was published in 1972, many of the performance comparisons from the Princeton Robustness Study have been largely superceded by the *biweight* M-estimate as an estimator of location (Gross, 1976, 1977). The biweight can be computed as an iteratively reweighted sample mean, where the weights w_i are defined in terms of the estimate obtained in the previous iteration, T_{n-1}:

$$
\begin{aligned}
T_n &= \sum_{i=1}^{n} w_i x_i \Big/ \sum_{i=1}^{n} x_i,\\
w_i &= w(u_i), \qquad u_i = (x_i - T_{n-1})/(cS),\\
w(u) &= (1-u^2)^2 I_{[-1,1]}(u),
\end{aligned}
$$

where c is a "tuning constant" (usually in the range 4–6) and S is an estimate of scale (usually made to be unbiased if the x_i's came from a Gaussian population). Beaton and Tukey (1974, pages 151–152) offered a rationale for the name *biweight*: "the 'bi-' referring to the outer exponent (whose value ensures continuity for both $w(u)$ and $w'(u)$)." The biweight has been shown to be highly efficient in many diverse contexts, including robust analysis of variance, time series and even in control charts used to monitor product quality.

6. ROBUSTNESS IN TIME SERIES

Brillinger (2002a) cited many instances where Tukey writes of robustness considerations in time series analysis. He further noted analogues between common robustness themes and time series procedures frequently advocated by Tukey. For example, outlier removal in the analysis of simple batches is analogous to detrending and point deletion in time series; likewise, the analogue to weighting is tapering and data windowing to minimize end effects, and detection of pairs of outliers is analogous to the detection of echoes in time series (Bogert, Healy and Tukey, 1963). In 1965, Tukey addressed the Institute of Geophysics and Planetary Geophysics in La Jolla and cited three areas of future importance to these scientists: "(1) Spectrum analysis, in the broadest sense, ... (2) Wise expression of data for analysis, ... (3) Modification of our techniques of summarization and analysis to deal with the case (almost universal, especially in geophysics, even after the observations have been expressed wisely) where fluctuations and errors have a distribution whose shape is far more straggling than that of the magic bell-shaped curve of Gauss and Laplace" (Tukey, 1965, page 660). With respect to these three issues, Tukey wrote:

> There are situations in which one may want to use nonlinear techniques to answer questions which are about linear phenomena and which, at least naively, would seem to be most naturally answered by linear procedures of analysis. In one situation—the analysis of time series for echoes—there is at least a moderate amount of evidence that this is a good thing to do. [*Tukey*, 1965, *page* 666.]
>
> If—often by so simple a change as taking square roots, or logarithms—we can eliminate most of these nonlinearities by a more appropriate and useful choice, perhaps guided by the bispectrum of a trial expression, we will be able to see much more deeply into these phenomena these time series describe. [*Tukey*, 1965, *page* 667.]
>
> If we must be realistic... if we are to get anywhere near the most out of our data, then we must face up to the possibility... that the distribution of our observations will often be one for which minimizing a quadratic is a very bad choice. And it requires only almost imperceptible modifications of the shape of a Gaussian distribution to make the use of the arithmetic mean a slightly inferior way to summarize the typical-value behavior of a sample, and to make its sum-of-squared deviations an almost unbearably poor summary of its spread (Tukey, 1960). The price of thinking more generally and more usefully here, of acting more realistically, will be that our procedures will become more nonlinear, and almost certainly iterative, and that, in particular, the procedures for assessing the quality of the result will have to be at least somewhat more complicated than evaluating a fixed quadratic function. [*Tukey*, 1965, *pages* 669–670.]

He repeated his emphasis on the need for robust methods in a paper on spectrum analysis designed for an advanced seminar at Madison in 1966: "It seems likely that the use of trimmed means, trimmed sums of squares, and trimmed sums of squared differences will prove useful in such a standard preprocessor [preparatory prewhitening]" (Tukey, 1968, page 828). Twelve years later, in "Can we predict where 'time series' should go next?" (Tukey, 1980), robustness was clearly on his mind:

> Both detecting and dealing with outliers are likely to need interpolated values, or sizes, for comparison. Here interpolation must be done robustly/resistantly since we do not yet know which are outliers and which are not. [*Tukey*, 1980, *page* 961.]
>
> Another question is: "How should we insert robustness/resistance into the ARMA calculations?" If we confine our attention to the AR part, which is often enough for a satisfactory approximation to prewhitening, then we can, as one possibility, give up any analogs of lagged moments and go directly to a robust/resistant regression of $y(t)$ on $y(t-1), y(t-2), \dots, y(t-k)$, as described by Denby and Martin (1979). [*Tukey*, 1980, *page* 970.]

In this paper, Tukey also discusses the impact and possible treatment of "holes," "isolates" and "outliers" in time series (Tukey, 1980, pages 966–967). In addition to such "exotic" values, the distribution of the data may be far from the conventionally assumed Gaussian distribution. Brillinger and Tukey (1985, pages 1020–1021) cite several instances where this occurs [in the following quotation, $X(t)$ denotes a random variable at time t, whose distribution serves as a probability model for a single real data value, say $x(t)$]:

> REAL DATA OFTEN FAIL to be Gaussian IN MANY WAYS.
>
> *Example 1*. Each realization can be thought of as a realization of Gaussian white noise with variance σ^2, but σ^2 varies from realization to realization.
>
> *Example 2*. Each realization is of the form $X(t) = \alpha\cos(\omega t + \phi)$, with α, ω and ϕ all random, where ϕ is uniform on $[0, 2\pi]$, independent of α and ω, which follow some messy joint distribution.
>
> *Example 3*. $X(t)$ is always either $+1$ or -1.
>
> *Example 4*. $X(t)$ is the sum of a realization of some fixed Gaussian colored noise and a peppering of random (say Poisson) values which affect only a small fraction of the observations (the contributions of this component are elsewhere zero).
>
> *Example 5*. $X^*(t) = X(t) + c[X(t)]^3$ where $X(t)$ is a realization of some fixed Gaussian colored noise.
>
> *Example 6*. $X^*(t) = [10 + \cos(\nu t + \phi)] \cdot X(t)$ where $X(t)$ is a realization of some fixed Gaussian colored noise, ν is fixed, and ϕ is a uniformly distributed random variate.
> It is easy to be—or become—non-Gaussian; it is hard to be—or remain—Gaussian. (Narrow-band filtration is an exception.)

In "Nonlinear (nonsuperposable) methods for smoothing data" (Tukey, 1974), Tukey discussed the mathematics, practice and implications of nonlinear smoothing, largely in the form of running medians. Tukey believed strongly in the use of 3R [running medians of length 3, repeated until convergence, which Arce and Gallagher (1988) and Wilson (1989) proved do indeed converge to "root signals," and for which Bryc and Peligrad (1992) proved a sort of central limit theorem for the smoothed output] and, more generally, in median smoothers, having devoted to them four chapters in *Exploratory Data Analysis* (*EDA*) (Tukey, 1977, Chapters 7, 7+, 8, 16) and two sections in "the green book" (Mosteller and Tukey, 1977, Sections 3F, 3G). In the last years of his life, he developed a further modification which he termed $3p$R, which involved a combination of smoothing, deletion and insertion, and which he hoped to include in *EDA2* (a revised version of *EDA*). The electrical engineering community later developed the idea of smoothing by medians into "separable median filters" (row medians, followed by column medians) for image processing (Narendra, 1981; Yang and Huang, 1981; Butz, 1988; Park and Lee, 1989). Brillinger (2002b) describes in greater depth Tukey's profound and long-lasting contributions to time series.

7. TRANSFORMATIONS

Because Tukey had analyzed so much data of so many different forms, he had direct experience with the various ways in which standard assumptions (e.g., normal distribution, homogeneous variances, linearity) do not hold. This experience led him to believe strongly in the value of transformations, as a way of "bending the data nearer the Procrustean bed of the assumptions underlying conventional analyses" (Tukey, 1957, page 603). Transformations had been discussed previously in the literature (e.g., Curtiss, 1943); many of these early articles considered their use for only certain types of data such as counts or proportions (e.g., Beall, 1942, Bartlett, 1947; Cochran, 1940) where distributions are often more Poisson or binomial than Gaussian. Freeman and Tukey (1950, page 607) reported on "an empirical study of a number of transformations, some intended for significance and confidence work and others for variance stabilization" specifically for data from these two distributions (Poisson and binomial). This paper led to Tukey's often-used transformations for Poisson counts, namely $\sqrt{4(\mathit{observed}) + 2} - \sqrt{4(\mathit{expected}) + 1}$ (using "$+1$" instead of "$+2$" in the first square root if $\mathit{observed} = 0$), because "clearly $\sqrt{x} + \sqrt{x+1}$ is best if small expectations are to be considered" (Freeman and Tukey, 1950, page 609), and $\sqrt{x} + \sqrt{x+1} \approx 2\sqrt{x + \frac{1}{2}} = \sqrt{4x+2}$, and, finally, because $E(\sqrt{X}) \le \sqrt{E(X)}$ for nonnegative random variables (Jensen's inequality).

A much longer and more detailed mathematical treatment of transformations appears in another one of his landmark papers, "On the comparative anatomy of transformations" (Tukey, 1957). Tukey recognized that a single transformation may not always achieve the desired purposes across the entire range of interest.

A classic example appears in *Exploratory Data Analysis*, where a single transformation fails to straighten the plot of U.S. population over the entire range 1800–1950: log(*population*) works well for Census counts between 1800 and 1870, while *population* (untransformed) works well from 1870 to 1950 (Tukey, 1977, Section 5D). Tukey later called such combinations of two transformations (here, log and identity) that are matched at a point "hybrid re-expressions," with the combination (square root and log, matched at the median) being called the "principal hybrid." [Emerson and Stoto (1983, Section 4E) discuss the procedure of matching any two transformations at a single point.]

Another approach to the problem raised by the fact that real data are often not Gaussian is to fit the data to a decidedly non-Gaussian distribution. For this purpose, Tukey proposed two classes of distributions, one for skewed distributions and the other for stretched-tailed (elongated) distributions, both relative to the Gaussian. Starting with a standard Gaussian random variable Z, the *g-distribution* is that of the random variable $Y_g(Z) = (e^{gZ} - 1)/g$; the parameter g indicates the amount and direction of skewness ($g = 0$ corresponds to the Gaussian). The random variable $Y_h(Z) = Ze^{hZ^2/2}$ has the *h-distribution*; larger values of h ($h > 0$) indicate longer, more stretched tails ($h = 0$ corresponds to the Gaussian). Combining both skewness and stretched tails, $Y_{gh}(Z) = e^{hZ^2/2}(e^{zZ} - 1)/g$. (In all three cases, location and scale can be incorporated by multiplying by a scale factor B and adding a location term A.) In his "Notes for Statistics 411," Tukey called the process of fitting these distributions to real data (i.e., fitting the parameters g, h, A, B) "*g–h technology*." Details and further results on the properties of these distributions, fitting them to real data, and applications of *g–h technology* can be found in Hoaglin (1985).

8. FURTHER WORK IN ROBUSTNESS

During the 1970s and 1980s, Tukey and others conducted much further work in robust methods; authors who have contributed to this literature are far too numerous to cite. Many of these ideas have appeared in *Exploratory Data Analysis* (Tukey, 1977) and *Data Analysis and Regression: A Second Course in Statistics* (Mosteller and Tukey, 1977), and in the three books edited by David Hoaglin, Fred Mosteller and John Tukey: *Understanding Robust and Exploratory Data Analysis* (1983), *Exploring Data Tables, Trends, and Shapes* (1985) and *Fundamentals of Exploratory Analysis of Variance* (1991), fondly known as *EDA*, *DAR*, *UREDA*, *EDTTS* and *FEAV*, respectively. Work motivated by problems with multivariate data led to methods of analysis and graphical displays (cf. Friedman and Tukey, 1974; Tukey and Tukey, 1981). Less widely available publications include many other procedures; one important example is *The Management of Weather Resources II*: *The Role of Statistics in Weather Resources Management* (Brillinger, Jones and Tukey, 1978), which discussed at length issues of the design and analysis of cloud-seeding experiments (many of which apply to experimental design more generally). Morgenthaler and Tukey (1991) considered the problem of finite-sample robustness in the book *Configural Polysampling*: *A Route to Practical Robustness* (see also Tukey, 1987), and robust multiple comparisons and graphical displays are important themes in Basford and Tukey (1998), *Graphical Analysis of Multiresponse Data*.

Many other robust-resistant procedures, and graphical displays designed for detecting departures from assumptions, are associated with Tukey's name: boxplot, stem-and-leaf diagram, robust-resistant line, confidence interval for the median, hanging rootogram, median polish, one degree of freedom for nonadditivity (Tukey, 1949), smoothing, twicing, 3R and many others that never made it into the formal published literature.

9. CONCLUSION AND REFLECTIONS

Robustness and exploratory data analysis permeated Tukey's work, from his earliest days as a statistician in the 1940s, continuing throughout his sixty-year, enormously productive career. He was ably assisted in this mission by several coauthors, including his long-time colleague and friend, Frederick Mosteller (e.g., Mosteller and Tukey, 1977). These concepts now influence everyone's work. Stem-and-leaf displays and boxplots are part of many grade school curricula. The incorporation of robust methods and consideration of departures from assumptions are routine today throughout much of the statistics literature. The focus on robustness and exploratory methods might have happened anyway, as computers became cheaper and more efficient at handling massive amounts of data—but Tukey certainly accelerated the development in these fields, allowing us to be somewhat more ready to "mine" data than we might otherwise have been able to do without him.

In 1982, Tukey ended his discussion on "The role of statistical graduate training" with the following prophesy: "We plan to influence what actually goes on, today and tomorrow. . . .We plan to help others in laying foundations for the future" (Tukey, 1982, page 889). And Tukey did exactly that. Lyle Jones found a quotation from James Thurber, who commented on the lasting influence of New Yorker columnist Harold Ross. The quotation applies to JWT as well:

> He is still all over the place for many of us, vitally stalking the corridors of our lives, disturbed and disturbing, fretting, stimulating, more evident in death than the living presence of ordinary men.

ACKNOWLEDGMENTS

My sincere thanks to David C. Hoaglin for his careful reading and excellent suggestions on this article, and also to the Editor and anonymous referee. (Any errors that remain are my responsibility.)

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.

ARCE, G. R. and GALLAGHER, N. C., JR. (1988). Stochastic analysis for the recursive median filter process. *IEEE Trans. Inform. Theory* **34** 669–679.

ARTHUR, S. P. (1979). Skew-stretched distributions and the t-statistic. Ph.D. dissertation, Dept. Statistics, Princeton Univ.

BARTLETT, M. S. (1947). The use of transformations. *Biometrics* **3** 39–52.

BASFORD, K. and TUKEY, J. W. (1998). *Graphical Analysis of Multiresponse Data*. Chapman and Hall, Boca Raton, FL.

BEALL, G. (1942). The transformation of data from entomological field experiments so that the analysis of variance becomes applicable. *Biometrika* **32** 243–262.

BEATON, A. E. and TUKEY, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics* **16** 147–185.

BOGERT, B. P., HEALY, M. J. R. and TUKEY, J. W. (1963). The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe-cracking. In *Proc. Symposium on Time Series Analysis* (M. Rosenblatt, ed.) 209–243. Wiley, New York. [Reprinted in *The Collected Works of John W. Tukey I. Time Series*: *1949–1964* (D. R. Brillinger, ed.) 455–493. Wadsworth, Monterey, CA (1984).]

BOX, G. E. P. (1953). Non-normality and tests on variances. *Biometrika* **40** 318–335.

BOX, G. E. P. and ANDERSEN, S. L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption (with discussion). *J. Roy. Statist. Soc. Ser. B* **17** 1–34.

BRILLINGER, D. R. (2002a). John Tukey and "troubled" time series. Presented at the 2002 International Conference on Robust Statistics, Vancouver, Canada, May 13, 2002.

BRILLINGER, D. R. (2002b). John W. Tukey's work on time series and spectrum analysis. *Ann. Statist.* **30** 1595–1618.

BRILLINGER, D. R., JONES, L. V. and TUKEY, J. W. (1978). *The Management of Weather Resources II*: *The Role of Statistics in Weather Resources Management*. U.S. Government Printing Office, Washington.

BRILLINGER, D. R. and TUKEY, J. W. (1985). Spectrum analysis in the presence of noise: Some issues and examples. In *The Collected Works of John W. Tukey II. Time Series*: *1965–1984* (D. R. Brillinger, ed.) 1001–1141. Wadsworth, Monterey, CA.

BRYC, W. and PELIGRAD, M. (1992). The central limit theorem for Tukey's 3R smoother. *Statist. Probab. Lett.* **13** 29–37.

BUTZ, A. R. (1988). Rank order smoothers with two-dimensional data models. *J. Math. Anal. Appl.* **136** 529–551.

COCHRAN, W. G. (1940). The analysis of variance when experimental errors follow the Poisson or binomial laws. *Ann. Math. Statist.* **11** 335–347.

CURTISS, J. H. (1943). On transformations used in the analysis of variance. *Ann. Math. Statist.* **14** 107–122.

DAVID, H. A. (2002). Personal communication (April 10, 2002).

DENBY, L. M. and MARTIN, R. D. (1979). Robust estimation of the first-order autoregressive parameter. *J. Amer. Statist. Assoc.* **74** 140–146.

DIXON, W. J. (1960). Simplified estimation from censored normal samples. *Ann. Math. Statist.* **31** 385–391.

DIXON, W. J. and TUKEY, J. W. (1968). Approximate behavior of the distribution of Winsorized t (trimming/Winsorization 2). *Technometrics* **10** 83–98.

EFRON, B. (1982). *The Jackknife*, *the Bootstrap*, *and Other Resampling Plans*. SIAM, Philadelphia.

EFRON, B. (1992). Jackknife-after-bootstrap standard errors and influence functions (with discussion). *J. Roy. Statist. Soc. Ser. B* **54** 83–127.

EFRON, B. and TIBSHIRANI, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

EMERSON, J. and STOTO, M. (1983). Transforming data. In *Understanding Robust and Exploratory Data Analysis* (D. C. Hoaglin, F. Mosteller and J. W. Tukey, eds.) 97–128. Wiley, New York.

FERNHOLZ, L. T. and MORGENTHALER, S. (2000). A conversation with John W. Tukey and Elizabeth Tukey. *Statist. Sci.* **15** 79–94.

FERNHOLZ, L. T., MORGENTHALER, S. and TUKEY, J. W. (2004). An outlier detection method based on the multihalver. *J. Statist. Plann. Inference*. To appear.

FREEMAN, M. F. and TUKEY, J. W. (1950). Transformations related to the angular and the square root. *Ann. Math. Statist.* **21** 607–611. [Reprinted in *The Collected Works of John W. Tukey VI. More Mathematical*: *1938–1984* (C. L. Mallows, ed.) 149–155. Wadsworth, Monterey, CA (1990).]

FRIEDMAN, J. H. and TUKEY, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.* **C-23** 881–890.

GROSS, A. M. (1976). Confidence interval robustness with long-tailed symmetric distributions. *J. Amer. Statist. Assoc.* **71** 409–416.

GROSS, A. M. (1977). Confidence intervals for bisquare regression estimates. *J. Amer. Statist. Assoc.* **72** 341–354.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

HASTINGS, C., MOSTELLER, F., TUKEY, J. W. and WINSOR, C. P. (1947). Low moments for small samples: A comparative study of order statistics. *Ann. Math. Statist.* **18** 361–426.

HOAGLIN, D. C. (1985). Summarizing shape numerically: The g- and h-distributions. In *Exploring Data Tables, Trends, and Shapes* (D. C. Hoaglin, F. Mosteller and J. W. Tukey, eds.) 461–514. Wiley, New York.

HOAGLIN, D. C. (2003). John W. Tukey and data analysis. *Statist. Sci.* **18** 311–318.

HOAGLIN, D. C., MOSTELLER, F. and TUKEY, J. W., eds. (1983). *Understanding Robust and Exploratory Data Analysis*. Wiley, New York.

HOAGLIN, D. C., MOSTELLER, F. and TUKEY, J. W., eds. (1985). *Exploring Data Tables, Trends, and Shapes*. Wiley, New York.

HOAGLIN, D. C., MOSTELLER, F. and TUKEY, J. W., eds. (1991). *Fundamentals of Exploratory Analysis of Variance*. Wiley, New York.

HORN, P. S. (1983). Some easy t statistics. *J. Amer. Statist. Assoc.* **78** 930–936.

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

JONES, H. L. (1956). Investigating the properties of a sample mean by employing random subsample means. *J. Amer. Statist. Assoc.* **51** 54–83.

MCLAUGHLIN, D. H. and TUKEY, J. W. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization. *Sankhyā Ser. A* **25** 331–352.

MORGENTHALER, S. and TUKEY, J. W., eds. (1991). *Configural Polysampling: A Route to Practical Robustness*. Wiley, 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. [Reprinted in *The Collected Works of John W. Tukey IV. Philosophy and Principles of Data Analysis*: *1965–1986* (L. V. Jones, ed.) 601–720. Wadsworth, Belmont, CA (1986).]

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

NARENDRA, P. M. (1981). A separable median filter for image noise smoothing. *IEEE Trans. Pattern Anal. Machine Intelligence* **3** 20–29.

PARK, S. Y. and LEE, Y. H. (1989). Double smoothing of images using median and Wiener filters. *IEEE Trans. Acoustics, Speech and Signal Processing* **37** 943–946.

PEARSON, E. S. (1929). The distribution of frequency constants in small samples from non-normal symmetrical and skew populations (2nd paper: The distribution of 'Student's z). *Biometrika* **21** 259–286. [Cited in Tukey (1948a), page 206.]

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

ROGERS, W. H. and TUKEY, J. W. (1972). Understanding some long-tailed symmetric distributions. *Statist. Neerlandica* **26** 211–226.

SCHEFFÉ, H. and TUKEY, J. W. (1944). A formula for sample sizes for population tolerance limits. *Ann. Math. Statist.* **15** 217.

SCHEFFÉ, H. and TUKEY, J. W. (1945). Non-parametric estimation. I. Validation of order statistics. *Ann. Math. Statist.* **16** 187–192.

SCOTT, D. W. (2001). Parametric statistical modeling by minimum integrated square error. *Technometrics* **43** 274–285.

SIMON, G. (1976). Computer simulation swindles, with applications to estimates of location and dispersion. *Appl. Statist.* **25** 266–274.

TUKEY, J. W. (1946). An inequality for deviations from medians. *Ann. Math. Statist.* **17** 75–78. [See also *The Collected Works of John W. Tukey VI. More Mathematical*: *1938–1984* (C. L. Mallows, ed.) lxviii. Wadsworth, Monterey, CA (1990).]

TUKEY, J. W. (1948a). Some elementary problems of importance to small sample practice. *Human Biology* **20** 205–214.

TUKEY, J. W. (1948b). Approximate weights. *Ann. Math. Statist.* **19** 91–92.

TUKEY, J. W. (1949). One degree of freedom for non-additivity. *Biometrics* **5** 232–242. [Reprinted in *The Collected Works of John W. Tukey VII. Factorial and ANOVA*: *1949–1962* (D. R. Cox, ed.) 1–13. Wadsworth, Monterey, CA (1992).]

TUKEY, J. W. (1957). On the comparative anatomy of transformations. *Ann. Math. Statist.* **28** 602–632. [Reprinted in *The Collected Works of John W. Tukey VI. More Mathematical*: *1938–1984* (C. L. Mallows, ed.) 167–209. Wadsworth, Monterey, CA (1990).]

TUKEY, J. W. (1958). Bias and confidence in not-quite large samples [abstract]. *Ann. Math. Statist.* **29** 614. [Reprinted in *The Collected Works of John W. Tukey VI. More Mathematical*: *1938–1984* (C. L. Mallows, ed.) 391. Wadsworth, Monterey, CA (1990).]

TUKEY, J. W. (1959). A quick, compact two-sample test to Duckworth's specifications. *Technometrics* **1** 31–48.

TUKEY, J. W. (1960a). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics*: *Essays in Honor of Harold Hotelling* (I. Olkin et al., eds.) 448–485. Stanford Univ. Press.

TUKEY, J. W. (1960b). Discussion of Anscombe and Daniel papers. *Technometrics* **2** 159–163.

TUKEY, J. W. (1961). Discussion, emphasizing the connection between analysis of variance and spectrum analysis. *Technometrics* **3** 191–219. [Reprinted in *The Collected Works of John W. Tukey I. Time Series*: *1949–1964* (D. R. Brillinger, ed.) 397–435. Wadsworth, Belmont, CA (1984).]

TUKEY, J. W. (1962). The future of data analysis. *Ann. Math. Statist.* **33** 1–67, 812. [Reprinted in *The Collected Works of John W. Tukey III. Philosophy and Principles of Data Analysis*: *1949–1964* (L. V. Jones, ed.) 391–484. Wadsworth, Belmont, CA (1986).]

TUKEY, J. W. (1965). Data analysis and the frontiers of geophysics. *Science* **148** 1283–1289. [Reprinted in *The Collected Works of John W. Tukey II. Time Series*: *1965–1984* (D. R. Brillinger, ed.) 659–675. Wadsworth, Belmont, CA (1984).]

TUKEY, J. W. (1968). An introduction to the calculations of numerical spectrum analysis. In *Spectral Analysis of Time Series* (B. Harris, ed.) 25–46. Wiley, New York. [Reprinted in *The Collected Works of John W. Tukey II. Time Series*: *1965–1984* (D. R. Brillinger, ed.) 811–835. Wadsworth, Belmont, CA (1984).]

TUKEY, J. W. (1974). Nonlinear (nonsuperposable) methods for smoothing data. In *The Collected Works of John W. Tukey II. Time Series*: *1965–1984* (D. R. Brillinger, ed.) 837–855. Wadsworth, Belmont, CA (1984).

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

TUKEY, J. W. (1979). Comment on "Nonparametric statistical data modeling," by E. Parzen. *J. Amer. Statist. Assoc.* **74** 121–122. [Reprinted in *The Collected Works of John W. Tukey IV. Philosophy and Principles of Data Analysis*: *1965–1986* (L. V. Jones, ed.) 805–810. Wadsworth, Belmont, CA (1986).]

TUKEY, J. W. (1980). Can we predict where "time series" should go next? In *Directions in Time Series* (D. R. Brillinger and G. C. Tiao, eds.) 1–31. IMS, Hayward, CA. [Reprinted in *The Collected Works of John W. Tukey II. Time Series*: *1965–1984* (D. R. Brillinger, ed.) 941–980. Wadsworth, Monterey, CA (1984).]

TUKEY, J. W. (1982). Discussion of "The role of statistical graduate training." In *Teaching of Statistics and Statistical Consulting* (J. S. Rustagi and D. A. Wolfe, eds.) 379–389. Academic Press, New York. [Reprinted in *The Collected Works of John W. Tukey IV. Philosophy and Principles of Data Analysis*: *1965–1986* (L. V. Jones, ed.) 881–889. Wadsworth, Belmont, CA (1986).]

TUKEY, J. W. (1987). Configural polysampling [The 1985 John von Neumann Lecture]. *SIAM Rev.* **29** 1–20.

TUKEY, J. W. (1990). Foreword. In *The Collected Works of John W. Tukey VI. More Mathematical*: *1938–1984* (C. L. Mallows, ed.) xlvii–l. Wadsworth, Monterey, CA.

TUKEY, P. A. and TUKEY, J. W. (1981). Graphic display of data sets in three or more dimensions. In *Interpreting Multivariate Data* (V. Barnett, ed.) 189–275. Wiley, Chichester. [Reprinted in *The Collected Works of John W. Tukey V. Graphics*: *1965–1985* (W. S. Cleveland, ed.) 189–288. Wadsworth, Belmont, CA (1988).]

WILSON, S. J. (1989). Convergence of iterated median rules. *Computer Vision, Graphics and Image Processing* **47** 105–110.

WINSOR, C. P. (1948). Factorial analysis of a multiple dichotomy. *Human Biology* **20** 195–204.

YANG, C. J. and HUANG, T. S. (1981). The effect of median filtering on edge location estimation. *Computer Graphics and Image Processing* **15** 224–245.

YUEN, K. K. (1974). The two-sample trimmed t for unequal population variances. *Biometrika* **61** 165–169.

YUEN, K. K. and DIXON, W. J. (1973). The approximate behaviour and performance of the two-sample trimmed t. *Biometrika* **60** 369–374.