John Tukey (1915-2000): Deconstructing Statistics[1]

James R. Thompson
Department of Statistics
Rice University, Houston, Texas 77251-1892
thomp@rice.edu

Abstract

John Tukey ushered in a new age of statistics. He taught us to question the fundamental assumptions of Fisher. He was the first significant postmodern statistician. John gave us a new paradigm of data analysis. He brought statistics into the age of computer visualization as an alternative to model based inference. He created a revolution whose ultimate consequences are still unclear.

Keywords: Exploratory Data Analysis, Graphics, Robustness.

1 Introduction

G.K Chesterton observed that there was a fundamental difference between the Anglo-Saxon and European (i.e., French) modalities of model-based action. The Europeans would build up a logical structure, brick on brick, to a point where conclusions, possibly dreadful ones, became clear and then act upon them. The Anglo-Saxons would go through the same process but then, before implementing the conclusions, would subject them to "common sense." "Yes, yes," they would say. "These seem to be the conclusions, but they simply aren't consistent with what we know to be true. Better have a reexamination of the assumptions."

Such an attitude has a long tradition in British letters and science. We recall Samuel Johnson when looking at a carefully crafted argument for determinism, including the absence of free will, dismissing the argument with, "Why, Sir, we know the will is free." And again, when considering Bishop Berkeley's neoPlatonist rejection of objective reality, Johnson kicked a stone down the high street as a conclusive counter-example.

Sometimes, even the British fall captive to the consequences of flawed models. For example, the Reverend Thomas Malthus's model opined that human populations would grow exponentially until they hit the linear (in time) growth of food supplies, and then crash, and start the process all over again. The ruling powers in

[1]This research was supported in part by a grant (DAAD19-99-1-0150) from the Army Research Office (Durham.

England quickly started acting on the implications of this model. The Irish Potato Famine was one of the results. Arguing that it would do no good to feed the starving Irish peasants from the plentiful, Irish produced, grain supplies, as this would simply postpone the inevitable, the English calmly watched while over a million people starved. The Malthusian model led to the (inappropriately named) theory called Social Darwinism, in reaction to which Marxism, with its own devastating consequences, was a response.

John Tukey was perhaps the most important questioner of models in Twentieth Century science. His logical roots extend all the way back to the Franciscan *Prince of Nominalists* William of Ockham (1280-1349).

There is a danger in disdaining the consquences of reasonable models. The AIDS epidemic is a case in point. Strong model based arguments were made early on for closing the gay bathhouses. They remain open in the United States to this day. Now there seems to be evidence that the huge pool of HIV infectives in the United States, via cheap air travel, drives a non stand-alone epidemic in Europe. But these are all model based conclusions, and model based conclusions are not the fashion today. An anti-modeling philosophy fits in well with the temper of the times. An American president has answered the ancient question, "What is truth?" with "It depends on what the meaning of 'is' is." We live in a time of grayness: "Some things nearly so, others nearly not."

Modern statistics was started in 1662 by John Graunt (Graunt, 1662), a London haberdasher and late Royalist captain of horse, without academic credentials, who came up with the incredible notion of aggregating data by quantitative attribute as a means of seeing the forest instead of the trees. Instead of discarding his ideas as downmarket because of the lack of credentials of the discoverer, the "respectable" establishment of his time, in the person of William Petty, simply tried to co-opt them as their own discovery.

The "Victorian Enlightenment," in which the statistical community played an important part, received its initial impetus from the Darwinian revolution in biological science. Statisticians were iconoclasts, questioning assumptions, using data based modeling to attack assumptions based on long held views. Francis Galton and Karl Pearson were key players, with Fisher coming along later as the premier consolidator and innovator of his statistical age.

A characteristic of these bold Victorians was a belief in their models. The old models of Lemarque, say, were wrong. The new models of the Victorians, however, were, if not precisely correct, a giant step toward reality. We recall Galton's confidence (Galton, 1879) in the universality of the "normal distribution" (so-called because everything followed it):

> I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error." The law would have been personified and deified by the Greeks, if they had known of it. It reigns with serenity and in complete self-effacement amidst the wildest confusion. The huger the mob and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason.

Students who have had their grades on tests subjected to a "normal curve" are victims of blind faith in Galton's 110 year old "maxim." (There is no obvious pooling of attributes which causes the Central Limit Theorem to operate in this case.)

2 Enter John Tukey and EDA

An obvious scientific progression would be for Galton's maxim to serve as an hypothesis which would compete with some new maxim which would then evolve into an improvement on the notion of universal normality. Tukey questioned the normal assumptions. He gave counter-example after counter-example where the normality assumption was false. But he never presumed to replace it with any model based alternative.

The same can be said for just about everything in the Fisherian pantheon of statistical models. Tukey deconstructed them all. But he did not replace them with a new set of explicit models. He simply dispensed with the Aristotelian paradigm of a chain of ever better models and changed the terms of discourse to the construction of a set of techniques for data analysis which were apparently "model free." Those who attended one of John's shortcourses in Exploratory Data Analysis (generally team taught with a junior colleague, to whom John gave most of the good punch lines) have heard the mantra:

> Exploratory Data Analysis lets the data speak to you without the interference of models.

3 The Box Plot

Let us consider one of the main tools of EDA: the *schematic* or *box* plot. The rules for its construction are rather simple:

1. Find the median (M) of the data.

2. Find the lower quartile and lower quartile of the data (lower hinge (LH) and upper hinge (UH)).

3. Compute the Step length: S = 1.5 (UH - LH).

4. Determine the Lower and Upper Inner Fences according to LIF = LH - S and UIF = LH + S.

5. Determine the Lower and Upper Outer Fences according to LOF = LH - 2S and UOF = UH + 2S.

The use of the *schematic* plot essentially replaces the normal theory based one-dimensional hypothesis tests of Fisher. Let us consider a sampling of 30 income levels from a small town: 5600, 8700, 9200, 9900, 10100, 11200,13100,16100,19300, 23900, 25100, 25800, 28100, 31000, 31300,32400, 35800, 37600,40100,42800, 47600, 49300, 53600, 55600, 58700, 63900, 72500, 81600, 86400, 156400.

$$M = \frac{1}{2}[31300 + 32400] = 31850. \tag{1}$$

$$LH = \frac{1}{2}[13100 + 16100] = 14600. \tag{2}$$

Similarly, for the *upper hinge*, we have

$$UH = \frac{1}{2}[53600 + 55600] = 54600. \tag{3}$$

$$S = 1.5 \times (54600 - 14600) = 60000. \tag{4}$$

$$UIF = 54600 + 60000 = 1146000. \tag{5}$$

$$UOF = 54600 + 2 \times 60000 = 174600. \tag{6}$$

In the Tukey paradigm, an income value outside an inner fence would be suspicious, i.e., we would not be very confident that it really was from the mass of incomes considered. And a value that is outside an outer fence would almost certainly not be from the mass of incomes considered. We see, here, that only one income is outside an inner fence, and it is inside an upper fence, namely the income of 156,400. So, by the use of the schematic plot, we would say that it appeared that the individual with an income of 156,400 is not a part of the overall population considered. Our judgement is ambiguous, since the observation falls inside the upper outer fence. Such a rule, as given in *Exploratory Data Analysis*, is didactic and, unlike the Fisherian significance tests, apparently model free.

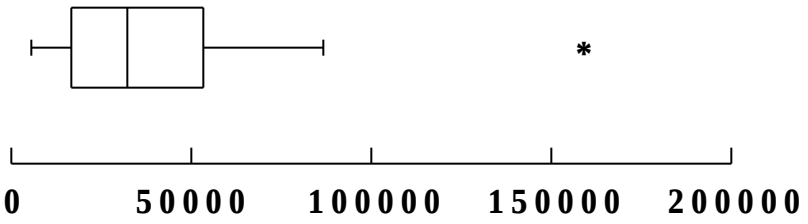

Figure 1. Box Plot of Income Data.

Let us note, however, some obvious assumptions. Firstly, the technique assumes that the data comes from a symmetrical distribution. (Of course, such symmetry might be approximately obtained from the raw data by an appropriate transformation.) Then, we need to notice some tacit assumptions underlying the placing of the

fences. If the underlying distribution is normal, the probability of an observation being outside an inner fence is $.007 \approx 1\%$. This is, interestingly, about the same one would obtain with a 1% significance test of parametric form. If the data is normal, the probability of an observation lying outside the outer fences is 2 in a million (a nice round number). It is hard to imagine that Tukey was not thinking in terms of the normal distribution as his standard. The inner fence and outer fence boundaries are not appropriate for distributions much tailier than the normal.[2] The technique based approach of the box plot, then, is very similar to what one would get if one used the normal assumption and specifically derived a likelihood ratio test, or the like.

If the schematic plot reduces essentially to a model based parametric procedure, what have we gained by its use? The answer would appear to be "graphical perception" rather than robustness. Moreover, the schematic plot does not correct for sample size. With a sample size of 30 from the same normal distribution, the probability that none of the observations will fall outside the inner fences is 81% rather than 99%. For 30 samples from the same normal distribution, the probability that all the observations will fall inside the outer fences is 99.994% rather than 99.9998%. Suppose the number of incomes we considered was 300 instead of 30. Then the probability of all the normal variates falling inside the inner fences is only 12.16%. The probability of all falling inside the outer fences is 99.94 %. The box plot is a wonderful desk accessory (and is installed as such in most statistical software packages), particularly when one is plotting data from two sources side by side, but it needs to be used with some care.

4 The 3RH Smooth

The work in nonparametric regression has been voluminous. Almost all has actually been within the context of Fisherian orthodoxy. Tukey attacked it *de novo*.

Consider a small company that makes van modifications transforming vans into campers. In Table 1 we show production figures on 30 consecutive days.

[2]For the Cauchy distribution, the probability of an observation falling outside an inner fence is 15.6%. The probability of an observation falling outside an outer fence is 9%.

Table 1. Production.	
1	133
2	155
3	199
4	193
5	201
6	235
7	185
8	160
9	161
10	182
11	149
12	110
13	72
14	101
15	60
16	42
17	15
18	44
19	60
20	86
21	50
22	40
23	34
24	40
25	33
26	67
27	73
28	57
29	85
30	90

We graph this information in Figure 2. The points we see are real production figures. There are no errors in them. Nevertheless, most people would naturally smooth them. Perhaps the figures are real, but the human observer wants them to be smooth, not rough. One could look upon such a tendency to want smoothness as being some sort of natural Platonic notion hardwired into the human brain. We can rationalize this tendency by saying that the real world has smooth productions contaminated by shocks such as sudden cancellations of orders, sickness of workers, etc. But the fact is that most of us would hate to make plans based on such a jumpy plot.

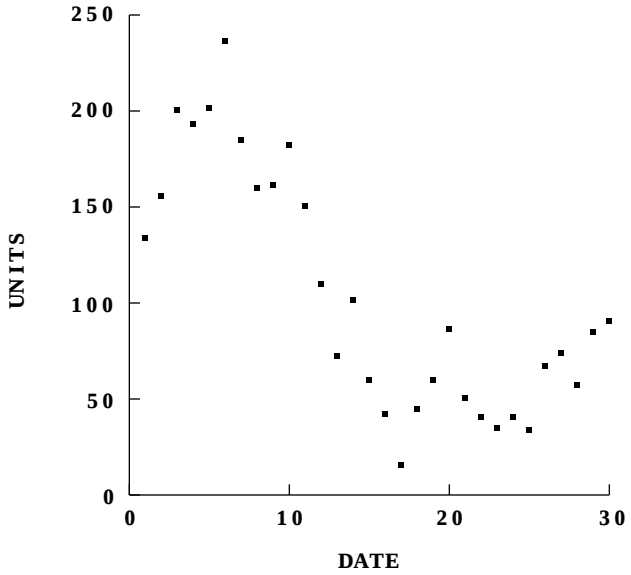

Figure 2. Daily Production Data.

The fact is that the world in which we live does tend to move rather smoothly in time. If there were so much turbulence that this was not the case, it would be hard to imagine how any sort of civilized society could ever have developed. So we ought not despise our apparently instinctive desire to see smooth curves rather than jumpy ones. Tukey achieved this by a running median (3R) smooth. Unlike the earlier hanning smooth (H) which replaced an observed value by a weighted (.25, .50, .25) average of it and its fellows, the 3R smooth will not drive the smooth to oversmoothed unreality when not checked. The 3R smooth is naturally self-limiting and requires no human manager. However, it is somewhat inclined to exhibit patches where one transitions from one locally flat function to another. So Tukey "polished" the 3R smooth with one or two hanning smooths. The result was an easy solution to the nonparametric regression problem, one that owed naught to normal theory or Fisherian orthodoxy. We show a 3RHH smooth of the production numbers in in Table 2 graphed in Figure 3.

Table 2. 3RHH Smooth.					
Day	Production	3	33=3R	3RH	3RHH
1	133	133	133	133	133
2	155	155	155	159	159
3	199	193	193	185	182
4	193	199	199	198	196
5	201	201	201	201	199
6	235	201	201	197	195
7	185	185	185	183	183
8	160	161	161	167	170
9	161	161	161	161	162
10	182	161	161	158	155
11	149	149	149	142	140
12	110	110	110	118	118
13	72	101	101	96	97
14	101	72	72	76	77
15	60	60	60	59	60
16	42	42	42	47	49
17	15	42	42	43	45
18	44	44	44	48	49
19	60	60	60	56	55
20	86	60	60	58	55
21	50	50	50	50	50
22	40	40	40	43	44
23	34	40	40	39	39
24	40	34	34	37	40
25	33	33	40	45	47
26	67	67	67	60	59
27	73	67	67	69	69
28	57	73	73	79	79
29	85	85	85	88	86
30	90	90	90	90	90

In the case of the box (schematic) plot, we have seen a technique of John Tukey which appears somewhat derivative of a Fisherian test of significance. The 3RHH smooth has nothing to do with Fisherian thinking. It is brand new with Tukey and is built on Gestaltic notions rather than anything dealing with normal theory.

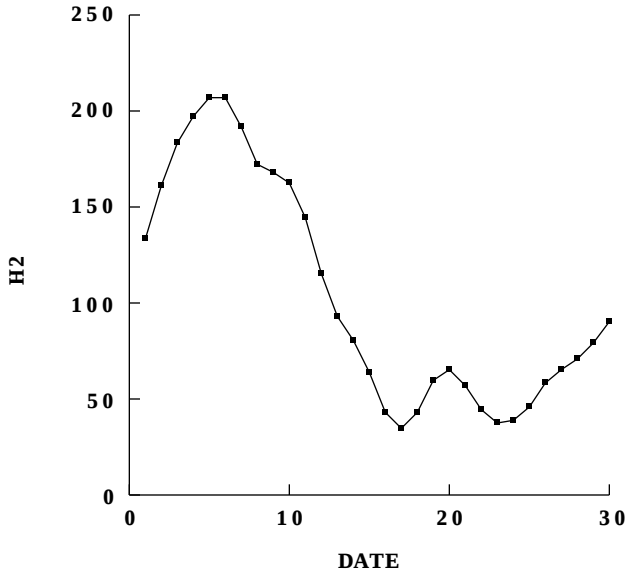

Figure 3. The 3RHH Smooth.

5 Conclusions

Most of the postmodern philosophers proclaim a new evangel and then wind up spouting old stuff from Nietzsche, Marx, Heidegger, Freud, etc. Tukey was a true revolutionary. Some of his material was, quite naturally, derivative from the work of Fisher, Pearson, and other statisticians from an earlier age. But most of the Tukey paradigm was new. In the case of smoothing, jackknifing, prewhitening, multiple comparisons, analysis of spectra, analysis of seismic data, citation indexing, projection-pursuit, Fast Fourier Transforming, data spinning, election forecasting, probably in most of the areas in which he worked, John Tukey was doing new things, things that nobody had done before. He deconstructed Fisher and Pearson and Gauss, but he then built his own structures.

What John did not do, unfortunately, was leave a roadmap of the models behind his paradigm. It can be said that W. Edwards Deming's legacy suffers from the same problem. Deming, like Tukey, was rather didactic, telling his disciples to do this or that because he, Ed Deming, had said to do it. But Deming was an Aristotelian, and, though he did not himself publish his modeling taxonomy, it is possible to discern it if one looks carefully. (Thompson and Koronacki (1992) have attempted to infer Deming's modeling structure.) Tukey is different. Nearly a pure nominalist, he spurns modeling and attacks problems almost sui generis. Yet there is a pattern. We can see some of John's implied axioms. Here are a few:

1. The eye expects adjacent pixels to be likely parts of a common whole.

2. The eye expects continuity (not quite the same as (1)).

3. As points move far apart, the human processor needs training to decide when points are no longer to be considered part of the common whole. Because of the ubiquity of situations where the Central Limit Theorem, in one form or another, applies, a natural benchmark is the normal distribution.

4. Symmetry reduces the complexity of data.

5. Symmetry essentially demands unimodality.

6. The only function which can be identified by the human eye is the straight line.

7. A point remains a point in any dimension.

Much more than this is required if one is to attempt to infer John's scientific philosophy. It might be an appropriate life's work for somebody, or for several somebodys. Though I have had a go at deciphering Deming, I despair of doing this for Tukey. And I am reasonably sure that if John were asked about the wisdom of somebody trying to come up with a Rosetta Stone for understanding the philosophy of John Tukey, he would laugh at the idea. And yet, if statistics is to survive, we have to come to grips with the fact that the statistical paradigm was changed by John Tukey in revolutionary ways, ways always brilliant, frequently destabilizing. He wanted to leave us a toolbox, and he did. But until we understand better the implications of the Tukey Revolution, we cannot possibly answer questions about the place of statistics in Twenty-First Century Science.

References

Brillinger, D.R., Fernholz, L.T. and Morgenthaler, S., eds. (1997), *The Practice of Data Analysis: Essays in Honor of John W. Tukey* Princeton: Princeton University Press, 40-45.

Cleveland, W.S.(editor in chief), *The Collected Works of John W. Tukey* (Eight volumes published from 1984 to 1994). Monterey, CA: Wadsworth Advanced Books and Software.

Galton, Francis (1889). *Natural Inheritance*. London: Macmillan & Company, 66.

Graunt, John (1662). *Natural and Political Observations on the Bills of Mortality*. London.

Thompson, J.R. (2000). "Is the United States Country Zero for the First-World AIDS Epidemic?." *The Journal of Theoretical Biology*, pp. 621-628.

Thompson, J.R. (1992), *Statistical Process Control for Quality Assurance*. London: Chapman % Hall.

Tukey, J.W. (1962), "The future of data analysis," *Annals of Mathematical Statistics*, 33, 1-67.

Tukey, J.W. (1973). *Citation Index,Index to Statistics and Probability, v.2* Los Altos, CA: R&D Press.

Tukey, J.W. (1977), *Exploratory Data Analysis*, Reading, MA: Addison Wesley Publishing Company.

Tukey, J.W. (1979), "Discussion of a Paper by Emanuel Parzen," *Journal of the American Statistical Association*, 74, 121-122.