Sunset Salvo

JOHN W. TUKEY*

My title is borrowed from Estill Green who used it when he retired as Executive Vice-President at Bell Laboratories about two decades ago. He was an executive, so he really retired. My salvo has a different flavor: research can and will go on; there will be a dawn tomorrow.

Let me first report on 44 years' observation of statistics and data analysis, the first 4 spent learning from Charlie Winsor.

Overall, things have become more diversified—more mathematical, more admittedly data analytic, more admittedly graphically displayed—and we can hope for continuing diversification; we need to struggle to keep this diversification balanced. In planning for balance we have to give special attention to one fact: The simplicity of writing theses in, and assessing the correctness of, mathematical statistics—like that of the mathematized sides of many other fields—will always ensure expansion of that corner. So our chief struggle has to be to emphasize all of the other directions.

HOW DO THE ARROWS POINT?

The essence of the arrows pointing from mathematical statistics toward data analysis includes these messages:

- Reduce dependence on assumptions, in particular by:

  1. using assumptions as leading cases, not truths (certainties are very rare)
  2. being explicit about using alternative assumptions (robustness; Chamberlain's definition, "Science is the holding of multiple working hypotheses!")
  3. when possible, using randomization to ensure validity—leaving to assumptions the task of helping with stringency (experiments can still be much safer)

- Increase the impact of results on consumers, in particular by:

  1. focusing on meaningful results (e.g., simple comparisons)
  2. using the most effective display techniques one can find to exhibit the results

- Take functional models more seriously than stochastic ones because:

  1. functional models are easier to verify
  2. as data accumulates, functional models become relatively more and more important
  3. statisticians have tended to overemphasize "their thing," namely stochastic models

- Expect to face up to more and more difficulties as the years pass by, in particular facing up to:

  1. lack of independence—particularly lack of detailed independence (individual numbers "close together" are often somewhat dependent)
  2. granularity, which is sometimes negligible but occasionally crucial
  3. lack of symmetry—a challenge we do not know how to meet: how can we then identify the *exact* question we want to answer?

- Pay more attention to the strategy of dealing with data, as compared with the tactics of carrying out a specified analysis:

  1. Doing this will be greatly aided by the needs arising in the growth of expert systems.
  2. Improved strategies will also need to be made available to those with limited hardware.

- Move from "all assumptions are right" toward "all assumptions are wrong."

SOME BRIEF DETAILS

Data analysis comes in various flavors; they will become more distinct.

We do not dare either give up exploratory data analysis (EDA) or make it our sole interest.

Jackknifery is more flexible than bootstrappery, in the sense that jackknifing by groups is more practiced than bootstrapping by groups—and it is often unreasonable to treat individual data elements that are "close together" as independent. Using the "proper error term" is often vital. Diversification within the use of internal error assessment comes naturally as diversified jackknifery. For the moment, jackknifery seems the most nearly realistic approach to assessing *many* of the sources of uncertainty—but remember that *no* fixed data-driven process can allow for systematic errors!

Graphical presentation can be made more forceful—and should be. In particular:

- We should not always plot points or even intervals.
- If, for instance, what we have learned is mainly what *cannot* be happening, then our pictures should emphasize what is ruled out—and neither what is left in nor what is most plausible. (Machine-produced graphics take away much of the need to save ink and drawing effort. We return to this point below.)

Most analyses of variance are used for exploratory purposes. We are just beginning to reach appropriately robust/resistant techniques. And we will then need to find appropriate confirmatory procedures.

Multiresponse techniques are in even less satisfactory situations.

The designed experiment is itself diversifying, in particular:

*John W. Tukey is Senior Research Statistician and Donner Professor of Science Emeritus, Fine Hall, Princeton University, Princeton, NJ 08544. This article is based on material presented or handed out at the 1985 Spring Symposium of the Northern New Jersey Chapter of ASA, May 6, 1985. It was prepared in part in connection with research at Princeton sponsored in part by Army Research Office (Durham), Grant DAAG29-82-K-0178.

• Taguchi is exploiting formalism as a route to understanding what appears to be going on—whether or not there is a sensible stochastic model. *F* tests serve him as attention directors, *not* as conclusion supporters, no matter what others may say.

• Computer reanalysis has made explicit limited randomization a desirable approach because:

1. We can use whatever summary statistic seems likely to be safest (valid—and of relatively high efficiency—in each of a variety of situations).
2. We can cope with the "overlap" of two or more analyses without the potential wastefulness of a Bonferroni correction.

STRENGTHENING OUR GRAPHIC DISPLAYS

Display is so important, and so taken for granted, that we need to give an example of how to do better. (The plots used come from a chapter by David Hoaglin and me in *Exploring Data Tables, Trends and Shapes*, edited by D. C. Hoaglin, F. Mosteller, and J. W. Tukey, which will be published by John Wiley and Sons. They refer to the classic measurements that indicated that radioactive disintegration appears to be a Poisson process.)

Figure 1 is an old-fashioned plot, designed to go with the pitch, "O, see how straight!" Good for propaganda, poor for careful study.

Figure 2 goes as far as even a small percentage of us are used to going. Instead of taking observed points as the likely corresponding long-run results (as in Fig. 1), it plots confidence intervals for each long-run result. Clearly better, but not good enough.

There are two ways in which we can make it better:

1. by flattening the picture so that we can make the intervals of uncertainty bigger and thus easier to work with;
2. by shifting emphasis from "what might be," inevitably truly fuzzy, to "what we know cannot be," which only has fuzzy edges.

Figure 3 shows what happens if we do both. We now have a good display for careful study. We can try fitting straight lines through the maze and see how many solid bars or open points they have to hit. (We return to further enhancement of this picture later.)

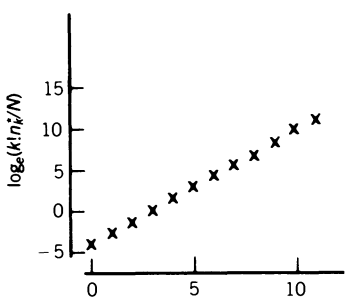

Figure 1. Emphasis on the Points Observed, as Our Best Estimates of the Answers (good as a general encouragement perhaps; bad as a part of careful analysis).

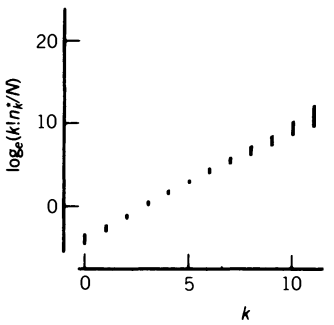

Figure 2. Emphasis on the Intervals Within Which the Answers Should Lie (better, particularly as part of careful analysis).

AVOIDING ASSUMPTIONS OF OMNICOMPETENCE

Here we have a major task: We need to avoid assuming that our viewers are omnicompetent. (It is usually not enough, and sometimes wasteful, to emphasize what was plotted.)

A diner in a restaurant can use a guide to gastronomy, but either a cookbook or an advanced course in cooking techniques would interfere with, rather than help, his enjoyment. Equally, a picture like Figure 3 can be greatly helped by an adequate sidebar commentary, like the following. (It is not enough to merely indicate what has been plotted on the two axes.)

COMMENTARY

An exact Poisson distribution should be a straight line in this plot. The solid bars, which tell where the long-run results are almost certain *not* to be, have been placed so that there are only about 4 chances in 100 of any single one covering a long-run result. Clearly, one can stick a straight line through without touching a solid bar. The open tapers have been so extended that there are only 46 chances in 100 that all of them will miss the corresponding long-run values. A straight line that hits only one open point is thus quite acceptable and can be done easily. Accordingly there is no reason why these data could not have come from an exact Poisson distribution.

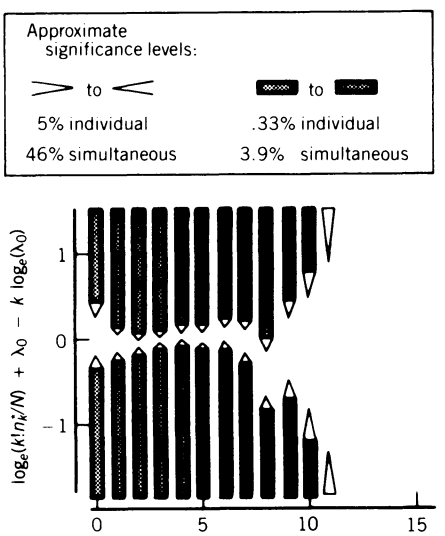

Figure 3. Emphasis on Where We Know the Answer Cannot Lie (best for careful analysis). See text for needed commentary.

To recoin and modernize a phrase: "A picture can be worth 1,000 words, but to be so, it may have to include 100 words."

SOME ROLES FOR "EXPERT SYSTEMS"

Many years ago a statistician whose name would be recognized if I revealed it spent a while at Princeton. One day he sat in my class when I talked about a practical thing to do. His reaction was, "How could you tell that to students? I would do it myself, but I would shut and lock the office door first?" This helps to illuminate the first of the two critical reasons why we have been unable to teach statistical data analysis (to the *large numbers* of people who need to use it):

1. We have been unwilling to think hard about what our real strategies (not just individualized tactics) have been or should be.
2. Even well-thought-out strategies would be too complicated to teach in the time available.

Our recent interest in expert systems has begun to change the first and promises to change the second:

1. One just cannot build an expert system without thinking through a strategy.
2. We should be able to teach initial use of each of several expert systems in a relatively short period of time. Using each well-planned system will then give continuing education—especially when the user repeatedly asks the system, "Why did it choose to do that?" After a while, some users will be ready for—nay will demand—more education, which we should by then be ready to furnish.

STRATEGIES FOR THE "POOR"

Many people will not have access to expert systems—because they use quite small computers or because they rely on batch computation. They too will need access to strategies. The reasonable way to serve them well will be with "strategy manuals," each probably tuned to a specific computing system.

MANAGING MATHEMATICS FOR THE GOOD OF A FIELD

Many—soon most—fields cannot get along without mathematics, usually in various forms, varying all the way from lowly computations of instances, one hopes selected to offer good guidance, to lofty flights of theorem proving. The usefulness of the latter usually suffers, perhaps necessarily, from constrained, unrealistic hypotheses. If our ultimate attention is on practice, we must

- recognize the incompleteness of both extreme approaches, as well as of those in between,
- do the best we can to balance our emphases in view of this incompleteness,
- use combined support—from diverse approaches—in those instances for which combined support is possible.

Putting more emphasis on theorem proving and less on computational approaches, as it would be too easy to be tempted to do, would also:

- Make it easier to evaluate "yes" versus "no," because this now tends to be only a matter of logical consistency
- Make it easier to avoid thinking about how well the framework—whether of instances studied or of "hypotheses"—fits the situation for which help is needed
- Make it easier to assign—and especially do—Ph.D. theses
- Make what we did much less useful

As a consequence, theorem proving is *seductive*—and its Lorelei voices can put us on the rocks.

There is usually a real need to do mathematical things at various levels of abstraction and an equally real tendency to do too much theorem proving and not enough numerical calculation that applies to more nearly real instances. In a world in which the price of calculation continues to decrease rapidly, but the price of theorem proving continues to hold steady or increase, elementary economics indicates that we ought to spend a larger and larger fraction of our time on calculation.

ASYMPTOTICS

A hard look at the practice and results of conventional asymptotic theory easily leads one to three points:

1. We ought *not* to expect it to be widely useful (since it refers to the ultimate oversimplification).
2. Rather often it has been useful (surprise!).
3. We have developed no basis for telling when it is likely to help and when it is not; the only general approach has been to do enough parallel work for finite samples and then see how asymptotics and finite samples join up.

A simple example of undue oversimplification is asymptotic robustness with fixed alternative shapes of distribution. If the sample size did get very large, we would not be taking any of the "fixed" alternatives seriously, since each one would already have been thoroughly contradicted by the data!

ANTIHISTAMINES AND ANTIHUBRISINES!

Our suffering sinuses are now frequently relieved by *antihistamines*. Our suffering philosophy—whether implicit or explicit—of data analysis, or of statistics, or of science and technology needs to be far more frequently relieved by *antihubrisines*.

To the Greeks *hubris* meant the kind of pride that would be punished by the gods. To statisticians, hubris should mean the kind of pride that fosters an inflated idea of one's powers and thereby keeps one from being more than marginally helpful to others. Hubris is the greatest danger that accompanies formal data analysis, including formalized statistical analysis. The feeling of "Give me (or more likely even, give my assistant) the data, and I will tell you what the real answer is!" is one we must all fight against again and again, and yet again.

Let me lay down a few basics, none of which is easy for all to accept, yet which all would be better for accepting:

1. The data may not contain the answer. The combination of some data and an aching desire for an answer does not

ensure that a reasonable answer can be extracted from a given body of data.

2. The data may not even contain an appearance of an answer, although we should look for appearances and then report them with adequate caution.

3. We must expect often to purvey appearances clearly labeled as such, rather than answers.

4. We ought usually to take our standard statistical (= confirmatory) task as assessing the minimum uncertainty to be assigned to the results found.

5. We can and should help our subject-matter–skilled clients with the assessment of systematic and other nonsampling errors to make the indicated uncertainty still larger; but we cannot often pretend to do the entire job ourselves.

6. Exploration, as purely a matter of seeking appearances, may need considerable aid from calculations, often rough, of minimum uncertainty. In the simplest cases, such rough calculations are built into our intuition. In more complicated ones, we must support our intuition with the results of numerical calculations.

Example: Additive Analysis of Factorial Patterns and Formal Factorial Analysis of Variance (ANOVA). Pigeonhole models for factorial ANOVA are so much more general than the classical Model I and Model II, with their Gaussian errors of everywhere the same variance! But it would be unrealistic to believe that even the pigeonhole models apply exactly to most instances of factorially patterned data. But if we think of the formal calculations of factorial ANOVA, classical or modernized, as in part a way of assessing minimum uncertainties, we can realize how useful such analysis has been and how useful it will continue to be. It need not be the main point of analysis of variance, however, as we now see.

Some of us believe 90% of all analyses of variance are exploratory in purpose and are happy to accept the usual ANOVA calculations as, in such cases, a sometimes useful supplement, reflecting minimum uncertainties. We will want to modernize the analysis of ANOVA as well as the model, but this will not get us beyond minimum uncertainties—although we expect the results to be even more helpful than in the past.

We would not have done as well with formal factorial analyses as we have if the analyses erected on probability-model foundations had not worked well in many cases in which these foundations were incomplete or absent. This applies to highly fractionalized patterns as well as to full factorials. When gross appearances are the "gold at the grass roots"—as appears, for instance, to be the case with Taguchi's lightly randomized (or unrandomized) factory-floor experiments—using the classical machinery of analysis of variance may be an easy route to effective exploration, whatever the cost in misperceptions and in difficulties of later reeducation. As the gold becomes harder to find, better experiments—particularly more randomization and more runs per experiment—will be needed and reeducation will be required, but it may well still be cheaper and quicker to do the early stages on an exploratory basis, whether admittedly so or not.

NUMERICAL ANTIHUBRISINES

Some of you will have heard my story of the chief chemist at an ordnance works who was so quantitative that his office was literally papered with graphs showing analytical results for different lots, but whose "chair sitting" estimates of standard deviation were about half of his "at wall looking" ones.

We owe it to ourselves, to our students, and to our clients to work hard on keeping a few simple numerical ideas in our minds and in theirs. Some very simple results about required sample sizes seem to be the natural beginning. (I thank Edwin Crow for pointing out undue optimism in my original versions.) Let us look at a few.

- How many observations, uncorrelated with variance σ^2, does it take to make the average length of a 95% confidence interval only 1σ? *Answer:* 18.
- How many to be sure that the *true* value is in a specifiable interval of length 1σ? *Answer:* More than 18; either by accepting a rough estimate of systematic error and increasing 18 greatly or by using some way of being sure that the systematic error is less than, say, $.1\sigma$. Doing the latter purely statistically with the same σ^2 often requires hundreds of measurements to assess the systematic errors.
- And if we wanted the same for $.2\sigma$? *Answer:* 400 measurements, and for the true value, either many more than 400 or so careful an assessment of systematic error as not to be doable in practice by measurement with variance σ^2.
- Measuring a change as a 95% confidence interval of average length only 2σ? *Answer:* Nine observations before and nine after.
- Measuring a difference equally well? *Answer:* 9 observations on each of the two, if we can pool variance estimates; 10 on each if we cannot. (Degrees of freedom are not all that important.)
- The same to $.2\sigma$? *Answer:* 800 before and 800 after (if we do not have to worry at all about systematic errors of comparison; otherwise many more).
- Measuring a σ^2 to one significant figure (taking this as, for instance, being reasonably sure for any $u > 0$, whether $2.5u$ or $3.5u$ is correct)? *Answer:* In the optimistic, Gaussian case, about 300 degrees of freedom. In more realistic cases, up to two or three times as many degrees of freedom. (Remember that Charlie Winsor's Principle says that no one ever has more than 100–200 degrees of freedom in practice, because by then heterogeneity of variance is beginning to show.)
- Measuring a correlation to a 95% confidence interval of average length .1 (an interval for r if r is small, for Fisher's z in general)? *Answer:* About 400 (x, y) pairs.

None of these is calming, many should raise blisters.

If we all repeated these examples to ourselves once a week, we *might* keep the following in mind:

- We may be able to get confidence intervals shortened to a not very small fraction of 1σ, but even this will be a lot of work.
- We will rarely, if ever, measure a σ^2 to one significant figure.

• Measuring a between-laboratories (or between-situations) σ^2 to one-half of a significant figure will take trials in 30 laboratories (or situations).

• Measuring a correlation coefficient, in terms of z, to $\pm .05$ is a lot of work; to $\pm .01$ is usually out of the question.

We all need to refresh our insight frequently by rehearsing these examples—there probably should be a more extended list for us to rehearse, too!

WHAT THEN SHOULD WE BE PROUD OF?

I have tried to make the following clear:

> It is dangerous hubris to think that we can solve all of the world's problems; but we can help many people in many fields, although what we can do is limited.

A story from 1937, when Ralph Boas was a postdoctoral student with Salomon Bochner helps make the last point. Ralph reported getting one of two answers whenever he told Bochner his latest results: "But zees is trivial!" or "But zees is impossible!"

With a reasonable amount of data, things of size 5σ are nearly "trivial" to find—anyone should be able to find them—whereas things of size $.05\sigma$ can be nearly impossible to find, once we face the presence of systematic error as well as those very nice errors whose effects come down like $\sigma/\sqrt{n}$.

Our pride should be greatest when what we can do is largest compared with what can be done without us—without regard to how simple or how complex are the ideas we use to do things better.

WHAT THEN SHOULD WE BE VERY PROUD OF?

• Questionably of getting more precision in our estimates, since it is probably only two powers of 10 from "But zees is trivial" to "But zees is impossible" and we are likely to be responsible for no more than one-half of a power of 10 of this gap.

• Probably of effective ways of dealing with responses to multiple factors (more detail shortly).

• Almost certainly of more effective ways of graphical presentation, especially those that are just beginning to appear.

To repeat and expand the second point, we should probably be very proud of having effective ways of dealing with responses to multiple factors—certainly in the classical analysis of variance with its extensions and modifications, provided we include data-driven choices among alternative breakdowns; probably in multiple regression, if we can succeed in bringing together almost all of the enhancements and insights that presently exist; and possibly in the analysis of multiway contingency tables, by the use of log-linear and other analyses.

CLOSE

In summary, then, my salvo has—or my salvoes have—been aimed at four important target areas:

• The diversity of data analysis—and its supporting statistics—has grown, and will grow further, and this growth will be hard to balance. Areas furthest from mathematical statistics will need emphasis most. It is important to know how the arrows point from mathematical statistics toward data analysis. (I have tried to list a number of the most important ones.)

• Both our own recognition of the strategy we use and the hope for the feasibility of teaching usable data analysis rest on the development of expert systems.

• Managing the use of mathematics will never be easy. Theorem proving is a seductive activity. In many situations asymptotics are a serious oversimplification, and we need to learn to know when asymptotics are likely to help.

• We all need to take a diverse collection of antihubrisines frequently: qualitative antihubrisines like "The data may not contain the answer!"; actual-role antihubrisines like "The mechanics of the analysis of variance can often be useful when the usual hypotheses have no connection with reality!"; and quantitative antihubrisines like "We're never going to measure a σ^2 to one significant figure!"

All of these target areas deserve continuing attention!

When I formally retire (at the end of June 1985) I do not plan to stop thinking or working. I plan to continue to provide both new techniques and new annoying-but-true statements.

[*Received August 1985.*]