ELSEVIER

Journal of Statistical Planning and Inference 57 (1997) 21–28

journal of statistical planning and inference

More honest foundations for data analysis

John Tukey[1]

Princeton University, Fine Hall, Washington Road, Princeton, NJ 08544-1000, USA

Received 1 July 1994; revised 1 May 1995

Abstract

Such foundations have to include *not* assuming that we always know what in fact we never know — the exact probability structure involved. We have to face up to bouquets of alternative challenges — of alternative probability structures that are relevant, yet extreme. And we have to focus on bouquets of alternative procedures, most, if not all, of which are not the results of formal optimisation. To select a procedure for use on a given sort of data set most carefully, we need to assess performance over combinations of procedures and challenges, often minimaxing over challenges. Such assessment will rarely be possible without simulation, perhaps of an analogous situation.

Selection of bouquets of challenges will involve experience, direct and extrapolated, as will selection of bouquets of procedures. Thus it will *not* be wrong for different experts of the greatest experience to choose different analyses of the same data. Moreover, the bouquets conventionally considered for any class of problem will evolve over time, thus changing preferred procedures. These are customary characteristics of other branches of engineering — data analysis, to be practical, has to be engineering, not science.

Model development, for real world description, *not* for inference, is a procedure of successive approximations where assumptions can be vital. Sound inference, on the other hand, demands facing alternative possibilities, as illustrated by diverse challenges.

AMS classification: 62A99, 62-07

Keywords: Challenges; Diverse sets of challenges; Evaluation of procedures; Extension of challenges; Worst relative performance; Styles of data analysis

1. Dangerous dichotomy

Many of us are aware that 'no data set is large enough to provide complete information about how it should be analyzed!'

[1] Prepared in part in connection with research at Princeton University, supported by the Army Research Office, Durham, DAAL03-91-G-0138.

PII S0378-3758(96)00032-8

Fewer of us have thought hard about the dangerous dichotomy of what to do about it:
(1) make a single assumption, possibly with estimable parameters, about a stochastic structure underlying the data,
(2) prepare a bouquet of diverse challenges, alternative stochastic structures (including distributional shapes) that might 'reasonably' underline the data.

Whether we go to a single assumption, or to a bouquet of challenges, we need next to consider, possibly implicitly, a collection of procedures, which might be applied to the data to give results, and to choose one (or possibly a few) to be actually applied to the data.

2. Facing challenges

In today's computerful world, more so tomorrow, we can use simulation to try out each of several procedures against each of several challenges. It is effortful rather than difficult to study procedures, or to compare them. We are not going to work with all reasonable challenges, the best we can hope for is to seek for diversity in the ways that seem likely to be important. As we learn more, we are almost sure to work with new challenges, and likely to work with more rather than fewer. However, the demands of diversity seem likely to grow less than proportionately to increase in computer speed, so our capacities seem likely to be stressed decreasingly.

3. Facing assumptions

It is usually impossible to study assumptions.

To study assumptions in the light of a given data set calls either for many such data sets which share a deeper, otherwise unspecified assumption, or for a much larger data set which shares such a real but unknown assumption. Whom do you know who has often had such opportunities? How many have had it even once?

If we go the route of single assumptions, they will almost always be unsupported assumptions. For me this is not a satisfactory foundation for a major human activity, data analysis.

The problem of finding data to test an assumption is exacerbated by the tendency of larger data sets to be less homogeneous; illustrated at the simplest level by Charlie Winsor's maxim 'no one ever had more than 100 or 200 degrees of freedom' usually because heterogeneity in large data sets keeps the stability of any mean square from being high.

4. Consequences of the challenges

If we go the route of challenges, we can expect that:

- at any one time, different analysts will feel that, for a specific data set, different challenges are important,

- any consensus about which challenges are important will change over time, so what we do in this decade, will not be the same as either what we did a few decades ago, or what we will do a few decades from now.

To some, this would seem giving up their *protective armor* of 'all reasonable analysts would agree' and their *safe refuge* in procedures that formally optimize given precise assumptions. Too bad, for the cracks in the armor are large, often allowing fatal wounds, and the safety of the refuge is illusory.

We live in a paradoxical world, where the only true safety, true though limited, comes from admitting both our uncertainty and the incompleteness with which we are able to meet it. Indeed, data analysis operates in the outside world, not in the world of mathematics.

In 1945, I was talking to Walther Mayer, then at the Institute of Advanced study. He had done war work, on the German side, in World War I and was surprised that I was staying on, part time, at Bell Laboratories. He had been so glad, as he put it, to go back where 'if he said a g_{ik} had such-and-such properties, it did!'. If data analysis is to work with real data, it cannot have that sort of mathematical certainty about its assumptions. We cannot change the real world by making assertions.

If we go the challenge route, working with several, hopefully well-selected, challenges where, in principle, we should consider a vast number, we admit to uncertainty and incompleteness but we are far safer than trying to huddle under a single assumption, usually an unsupported one.

If we can bear to face our uncertainties, and, unless we do, either data analysis or our role in data analysis will inevitably suffer, then it is time to think in a little more detail about the steps that have to be taken, explicitly or implicitly, in data analysis.

5. Choice of challenges

How should we, indeed how can we, reasonably choose challenges?

First, we should try for challenges that span the stochastic situation that best represents the origin and specificity of the data. Here 'span' implies both diversity and location near the edges of reasonability. We should begin by drawing on the ambient understanding of what possibilities need to be considered, and can be handled. Some of this understanding will depend on our insights on which challenges stress one or more popular procedures severely. Then we should examine the consistency of each challenge with the data before us, eliminating or modifying challenges for which the data are too unlikely. If we are lucky, this process may suggest new challenges for investigation.

6. The simplest example

In the simplest situations, the first step has evolved along the following path:

(1) the observations follow Gaussian distributions with known variances and means related by some structural model,

(2) as (1), but variances are in known ratios but of unknown size,
(3) as (2), but alternatively Gaussian or stretched-tail in shape of distribution (almost equivalently, that variances differ, probably substantially and in an unknown pattern),
(4) as (3), but also differ, alternatively in other ways (not yet formulated).
We will have to learn to regard level (2)'s Gaussian as quite inadequate in the sense that most of us were brought up to regard level (1)'s known variance as quite inadequate.

Given a large enough bouquet of challenges to begin with, the second step can involve either:
(1) dropping a challenge entirely, when the data are too discrepant with it, or
(2) shrinking a challenge inward, for example when a very stretched-tail challenge is replaced by a moderately stretched-tail challenge, because, of the two, only the latter is reasonably compatible with the data.

7. Choice of candidate procedures

The first step with procedures is to assemble a set of candidate procedures that deserve consideration. This set is naturally diverse, both in character and in source. Some procedures will appear because they are classical, like unweighted least squares (even though we are clear about its real fallibility). Others will appear because of repeated good experience with their use, such as iteratively reweighted least squares. Still others will appear because of simplicity and understandability, such as analysis of cell medians, or cell midmeans, for data in a factorial structure. And still others may be suggested by some of the less familiar challenges, perhaps via George Easton's (1991) adaptations of asymptotic expansion to robustness, which should apply to more general challenges.

8. Evaluation of procedures

To back one procedure, or perhaps a few procedures, in a well-supported way for use on our data set we need to ask about the 2-way tables of performance as a function of challenge and procedure. To construct such a table, we need a choice of performance measure and a substantial number of simulations. It is unlikely that we will get this amount of information for even every substantial data set, although it is clear that is what we might well have to do to be careful.

More often than not, we will borrow insight or choice from previous experience with at least rather similar situations. To borrow insight may mean using experience to pinpoint where difficulties are likely to be greatest, leaving us with many fewer challenges to investigate and evaluate. In simple situations, for instance, we may borrow insight that equally, or nominally, weighted least squares will be in greatest

trouble (a) with the challenge(s) where tails are most stretched and (b) with the challenge(s) where actual ratios of variances deviate most widely from the corresponding nominal ratios. To borrow choice might mean following the past by using a procedure often used as a natural compromise between Gaussian and stretched-tail challenges.

But suppose we have agreed upon a performance measure, and have obtained a two-way table of its values, for, say, 10 challenges and 20 procedures, how are we to choose one of the twenty? How do we judge the procedures, given any one challenge? Clearly by comparison, and by comparison with something that provides good performance against this specific challenge. The nature and properties of the performance measure will have to determine whether 'comparison' should involve 'difference' or 'ratio' (or possibly some other combination). The reference standard might be, for example,

- if we know it (only for the simplest challenges), the best that any procedure can do against the given challenge,
- the best any of the 20 procedures does for this challenge,
- the midmean of the four best performances (of the 20).

In any case we can think of our relative measure as 'how far this procedure falls short of the reference'. Occasionally, some comparisons may be opposite, and some properties may 'fall long', rather than fall short, by exceeding the reference for some challenge.

If the challenges have been well chosen, to be reasonable, diverse and near the edges of reasonability, then combining amounts of 'fall short' across challenges deserves emphasis on the worst, most simply on the worst relative performance for any challenge. After all, if the challenges are realistic, each, including the worst, is something that could be happening to the procedure, were we to apply it to the data set.

So long as amounts of shortfall are not large, which seems to be what we have found to date, choosing the procedure with the best, i.e. least, value of the worst relative performance seems quite appropriate. (We have never yet faced a situation where shortfalls are large and point in different directions. When we do, we may need to offer alternative procedures, and thus alternative results, depending upon which challenges are taken *more* seriously.)

This discussion corresponds to, in the (very simple) robust location case, looking at

$$\text{relative efficiency} = \frac{\text{variance for best procedure}}{\text{variance for this procedure}}$$

and

$$\text{polyefficiency} = \text{minimum of relative efficiencies (over challenges).}$$

Indeed, the whole approach we have been describing is clearly one generalization and philosophization of the 'finite alternatives' approach to robustness. It regards, properly I believe, what has been done in that arena, not as a compromise with practicality, but rather as an instance of the use of the highest principles.

9. Description of results

Now we have, say, one procedure, but still 10 challenges. This means, typically, one set of primary results, for example, estimates of locations, but, again typically, alternative secondary results, for example, estimated variances of estimates of location, which will often depend upon the challenge as well as upon the data. We can expect different secondary results for each challenge. How are we to put together these secondary results? The challenges were supposed to be (barely) reasonable, extreme examples of what might really be going on. This fact leaves me feeling forced to report the worst (e.g. the largest estimated variance) of these secondary results. (In cases where the procedure allows the challenge to have some small influences on primary results, we should need to take account of this too.)

10. Comments

The overall picture will seem unduly loose to many, but there seems to be no way to combine tightness and realism, and realism must win in the end. The process will appear to be one of seeking out challenges that made the result look worse, but this fact argues for reasonably diverse challenges as the way toward greater realism. (The antithesis, seeking out assumptions that make the results look better and better, is surely unacceptable).

I have pointed out earlier that the life of statistical procedures seems to be on a scale resembling 3 tridecades, 90 years from beginning to end. This leaves me unsurprised about major changes in practice being observed every few decades, changes such as from non-robust to robust, which is clearly taking decades.

Almost nothing is harder than verifying a stochastic assumption, something I find hard to believe has ever been done in even a single instance. We can hardly use unvalidated assumptions as foundations for an honest formulation of data analysis. Since we have no other kind of assumptions, we cannot found a honest account of data analysis on assumptions.

The 'loyal opposition' will argue that bouquets of challenges are just sets of unvalidated assumptions. In this, they will clearly have a point. But there is an appropriate reply. A good (for the time and subject) bouquet of challenges reflects our understanding, at the time and for the sort of data at hand, of what stochastic behavior is most challenging for that data's analysis. And the whole framework expects the inclusion of new challenges when, and if, we learn about stochastic structures that challenge analysis in new ways.

We propose to replace an approach with all its uncertainties swept under a small rug, by concealment in single unvalidated assumptions, by a more flexible and more parallelized approach that is less certain about the world, but which emphasizes dealing with those aspects of uncertainty with serious consequences by parallel challenges, calculation of relative performance, and emphasis on the worst that may reasonably happen (rather than on the best that conceivably might). To me this seems that only route toward a reasonable degree of facing the facts and a reasonable degree of admission that the millennium (not just the year 2000) has not yet come.

Most of us have seen substantial change in methods of data analysis, at least in restricted fields. Our younger colleagues can find this by reading literature of appropriate age. It would be a very poor inference to conclude that change is now finished, rather we ought to conclude that changes in the past imply changes in the future.

11. No linkage with style of analysis

The paradigm shift about the foundations of data analysis is in no way logically linked to the style of data analysis involved. The relative importance of some aspects may alter somewhat, but the combination of 'no single assumption' 'multiple challenges' has a place in all styles of data analysis, including rerandomization, exploratory data analysis, common or robust analyses of factorially arranged data.

The Platinum Standard of rerandomization, often double (two-step) randomization, where the probability aspects of the conclusions are completely validated by our randomness in assigning treatment to units, works for any choice of procedure. But it produces better (i.e. more stringent) results, if the procedure chosen does well in the face of the actual, stochastic situation inherent in the data generation. Picking a procedure is no longer a matter of validity of results, but it remains a matter of quality of results. And we can be surer of quality by guiding choice of procedure with a bouquet of challenges rather than with a single, usually one-sided, assumption.

Exploratory Data Analysis, at the other extreme, focuses on appearances (often primary results) rather than on conclusions or significance (all of which must also involve secondary results). From its earliest days, it has sought methods that are effective under diverse stochastic circumstances, without emphasis on the highest possible stringency. Indeed, it is truly procedure-based rather than challenge-based. One could try to do exploratory data analysis on the basis of standard single assumptions; single assumptions and exploration can be combined, but we know enough not to try this.

Intermediate styles, such as anova-related analysis of factorial arrays of responses, can also be done in either framework. Doing it in a single-assumption framework (e.g. Hoaglin et al., 1991) can be a substantial improvement over classical approaches, but doing it in a multiple-challenge framework (e.g. Seheult and Tukey, 1996) produces results that are often different and, once seen, hard to forget.

And so on, I believe, for all the other styles of data analysis.

The choice between single assumption and diverse challenges underlies data analysis as a whole. Diverse challenges offer a much more honest foundation for dealing with actual data.

12. Disclaimer

Some have taken my words about not trying as a crutch for inference to fit, or aribtrarily assume, a single stochastic model as words against fitting a stochastic model in any circumstances. This is a clear and dangerous error. There are many situations where we want the single best stochastic model we can get. My point is that *inference*, formal or informal, *is not one* of these situations. (More attention to what other stochastic model are consistent with what we know may well be worthwhile in non-inferential situations, but diversity there is usually far less important than for inference).

Others may believe I would trash analysis based on an unsupported Gaussian distribution. If such an analysis is all we have, we should probably use it, but hoist a red flag reading '*danger*: *this is no more than a preliminary procedure*', (Single challenges may continue to be the natural instances with which to begin, but only if their use cries out for facing diversity.)

13. Close

It is now almost *nine* decades since we began to stop treating estimated variances as if they were known and began to use Student's t, instead of its limiting Gaussian distribution, to evaluate means of smaller samples. It is now more than *three* decades since we began, for many ever so reluctantly, to stop treating assumed shapes of distribution as if they were known to be Gaussian, and started to use robust methods, at least as parallels to means and ordinary least squares.

It is high time to recognize such paradigm shifts as things to be anticipated, as steps forward (not backward), as part of a mechanism through which the results of data analysis become more trustworthy.

We need to use robust procedures much more widely. We need to ask: what further broadening of challenges would be most important? And then to ask: how can we start to meet these broader challenges?

References

Easton, G.S. (1991). Location compromise maximum likelihood estimators. In: Stephan Morgenthaler and John W. Tukey, Eds., *Configural Polysampling – A Route to Practical Robustness*. Wiley, New York, Chapter 11, 157–192.

Hoaglin, D.C., F. Mosteller, and J.W. Tukey, Eds. (1991). *Fundamentals of Exploratory Analysis of Variance*. Wiley, New York.

Seheult, A. and J.W. Tukey (1996). In preparation.