# Issues relevant to an honest account of data-based inference, partially in the light of Laurie Davies's paper

*John W. Tukey*[1]
*Princeton University, 408 Fine Hall, Washington Road, Princeton, NJ 08544-1000*

## Abstract

The changes in conceptual structure discussed in Davies 1993+ are important but incomplete. The present account discusses briefly a broader set of issues - - still likely to be incomplete. Just how far we shall ultimately have to go is uncertain, but we can see quite a lot of what needs to be thought about and built into our plans and procedures. After an introduction to procedure orientation, a number of issues are taken up in alphabetic order.

[1] Prepared in part in connection with research at Princeton University, supported by the Army Research Office, Durham, DAAL03-91-G-0138, and facilitated by the Alfred P. Sloan Foundation. July 27, 1993

# Contents

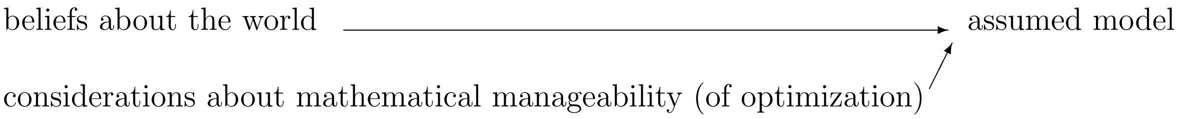
# Introduction to procedure orientation

Davies has gone a long way toward procedure orientation. Her discussion at the top of her page 29 is summarized by the sentence (material in [ ] added): "In other words models are chosen to produce [as properties of the procedures they 'prescribe'] desirable operational characteristics."

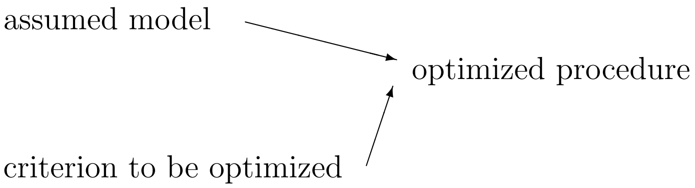
This is a long step forward, models are now valued for their (formal) consequences, rather than for their truth. This is good, but does not go far enough. Since the formal consequences are consequences of the truth of the model, once we have ceased to give a model's truth a special role, we cannot allow it to "prescribe" a procedure. What we really need to do is to choose a procedure, something we can be helped in by a knowledge of the behavior of alternative procedures when various models (= various challenges) apply, most helpfully models that are (a) bland (see below) or otherwise relatively trustworthy, (b) reasonable in themselves, and (c) compatible with the data.

In short, we need to change from assumption-orientation to procedure-orientation.

The classical paradigm ran something like:


beliefs about the world
assumed model
considerations about mathematical manageability (of optimization)

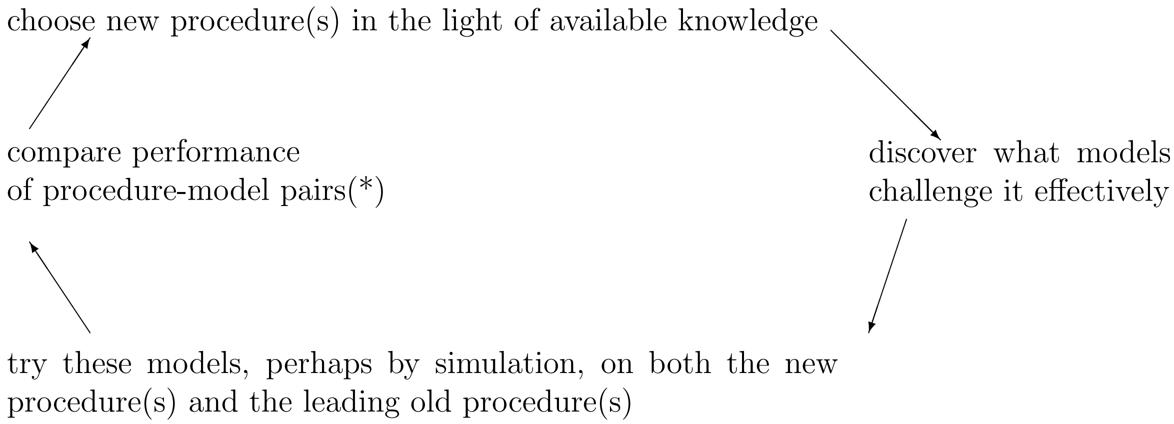
where, as Davies rightly emphasizes, the process involved in these first two arrows has been very inadequantely discussed, followed by



The solidity of the optimization was classically taken as legitimizing the unremovable fluidity of the choice as [sic] assumed model and of the criterion to be optimized.

Since the model was about assumed (revealed?) truth, one could try to get away with one model and with qualitative knowledge about optimization – what procedure optimized rather than how well it performed in contrast to alternatives.

Once we have become procedure-oriented, we expect a very different pattern of thought. Beginning somewhere, we enter a loop of exploration and improvement:



choose new procedure(s) in the light of available knowledge

discover what models challenge it effectively

try these models, perhaps by simulation, on both the new procedure(s) and the leading old procedure(s)

compare performance of procedure-model pairs(*)

At (*) we are likely to seek some sort of saddle-point, where deviations of model move performance in one direction, while deviations of procedure move it in the other.

Once we have taken simultation seriously, the minimum mathematical manageability we require is limited to:

- being able to apply procedures to data, and
- being able to simulate data from models

so that we no longer require an ability to carry out complex formal manipulations.

* framework *

We need to frame our thinking in terms of:

- questions, on which we hope the data - - and its analysis - - will shed light,

- targets - - one or a few numbers or pictures (or both)- - to be calculated from the data by the use of

- procedures - - well-specified computations leading from the data to values of numerical targets or examples of pictorial targets.

(The narrowness of targets reflects the limitations of human preception.)

# 1 A is for Approximation

Davies's emphasis on approximation is well-chosen and surprisingly novel. While these will undoubtedly be a place for much careful work in learning how to describe the concept - - and its applications - - in detail, it is clear that Davies has taken the decisive step by asserting that there must be a formal admission that adequate approximation, of one set of observable (or simulated) values by another set, needs to be treated as practical identity.

If, as is so convenient, we continue to use continuous models to describe - - or perhaps only to illuminate - - observed data, we should have to say that certain aspects of the data - - not typically, but unavoidably, including "Most (modelled) observations have irrational values!"- - are not to be used in relating conceptual (or simulated) samples to observed samples. Thought and debate as to just which aspects are to be denied legitimacy will be both necessary and valuable.

# 2 B is for Blandness

Davies emphasizes the pathological discontinuity of estimate behavior as a function of assumed model. In any neighborhood of a simple model (location, or location- and-scale, say) there are arbitrariy many (similarly simple) models with arbitrarily precise potentialities for estimation - - potentialities that can only be realized by very different procedures that [sic] those that realize the potentialities of the original simple model.

There are always nearby models that "describe the data" - - and are arbitrarily easier to estimate from. There may or may not be models that "describe the data" and are harder to estimate from.

A slightly pessimistic (or cynical) view of the world is that we can only trust inferences from models that are (locally) hard to estimate. Nothing that I know about the practicalities of data analysis offers any evidence that this view is too pessimistic. What we can expect, therefore, is that we ought to find trustworthy and helpful - - not as *prescribing* procedures, but as useful challenges to procedures, leading to useful illustrations of how procedures function - - those models that are (locally in the space of models) hard to estimate. Locally easy models are not to be trusted.

From one point of view, that we are putting forward is a two-part description of the meaning of the acronym

tinstaafl

well know to science-fiction readers, which stands for

THERE'S NO SUCH THING AS A FREE LUNCH,

the two parts being expressible in similar terms, but more closely relevant to our context as

NO ONE HAS EVER SHOWN THAT HE OR SHE HAD A FREE LUNCH

Here, of course, "FREE LUNCH" means "usefulness of a model that is locally easy to make inferences from".

Any very specific characteristic of a distribution is a potential handle by which easier estimation is possible. So we want models that are as *bland* as we know how to make them as our prime guides - - our prime illumination of the behavior of alternative procedures.

There is a sense in which the most useful interpretation of the Central Limit Theorem is not about how nearly Gaussian the distribution of estimates is likely to be, but rather as an indication of blandness of Gaussian models - - since they arise, in the limit, by the disappearance of any and every kind of distinctive behavior. As one of several, or even many, models useful in illuminating behavior of some procedure, then, Gaussian distributions probably do have a special role.

## * blandness in finite samples *

One outcome of the Princeton Robustness Study [1](Andrews et.al. 1972e) was the recognition of the non-blandness - - in small samples - - of Student's $t$ with very few degrees of freedom. The peakedness of $t_1$(and, to a lesser extent, of $t_2$) was found to be sufficient for an estimate of location, tuned to make use of this peakedness, to have 10's of % more efficacy than is possible for a model with the same (seriously challenging) tail behavior but smaller peakedness (for example the slash or $h_1$ distribution - - the distribution of a centered Gaussian deviate divided by an independent rectangular [0,1] denominator.)

There is thus a finite sample sense in which $h_1$ is much blander than $t_1$ - - and a finite sample sense in which we can reasonably ask if $h_1$ is bland enough.

Whether there will ever be detailed and rigorous formulations of blandness - - for either floating (asymptotic) $n$ or finite $n$ - - that will be directly useful seems presently to be less than certain. But that does not excuse us for not asking about blandness in every situation where we are trying to understand the behavior of a chosen procedure against a background of well-selected alternate models.

* blandness of asymptotics *

Davies makes much of the discontinuity of asymtotics in the topology of classical inference. Here [sic] remarks are accurate, her conclusions unhelpful. This is another" free lunch" issue. There could be asymptotic free lunches, but no one seems demonstrably to have ever had one.

If asymptotics are of any real value, it must be because they teach us something useful in finite samples. I wish I knew how to be sure when this happens.

But leaving this basic doubt aside for the present, classical asymptotics (like those cases where they do their best) should be read as saying:

1) Things might not be any better than so-and-so for large samples.
2) No one has a real example where they are better.
3) Thus they give both lower bounds and reasonable choices for an appropriate diffidence (= lack of confidence).

We need to do asymptotics based on bland models - - how lucky we are that the Gaussians is relatively bland.

## 3 C is for Challenges

How then do we seek for good procedures? We do *not* try to find a single model case that is as nearly truthful as we can stand (our tolerability or intolerability is usually from the point of view of mathematical manipulability). We do try to find a suite, so far relatively small, of models which we hope are effective in challenging the interesting procedures, while still compatible with the data.

We seek, not a unique truth but rather a diversity of challenges.

Because the term "model" is so tightly bound to ideas of truth, we will in our procedure-oriented accounts of data-analytic inference use the word "challenge" and the phrase "suite of challenges" instead. The idea is that the challenges in a suite will all - - or perhaps almost all - - be compatible with the actual data on which our inference is to be based. The most effective compatible challenges are almost certain to be among those where the compatibility is marginal. This means, in particular, that the suite will have to change somewhat for substantial changes in sample size.

For most really large data sets, indeed, it will probably be desirable to allow for more structure in the data then that of a simple random sample - - 10,000 observations, for example, might need to be divided into perhaps a couple of hundred groups of average size 50 in a way that reflects how the data was gathered - - and the challenges would need to reflect group-to-group differences. Among the corresponding challenges we would need to include challenges in which these groups differed in level, or spread, or both.

## 4 D is for Datesware - - specifically Consumer Datesware

Some people - - of course not including the present writer - - seem to feel that classical inference - - classical mathematical statistics has some central purpose other than helping to analyze data. But surely no one thinks that procedure-oriented choices among procedures has any non-data-analytic purpose.

If the aim procedure-oriented guidance is to aid in the analysis of data, those who have done the most to plan and/or conduct such choice have an obligation to assist in making the results of such choice useful to those who have data to analyze. This means making specific recommendations about which algorithms to use for which purposes under which conditions - - making all the data analytic choices needed to specify data-analytic software.

Neither the history of the last century, nor the present list of issues suggests finality of choice of procedure as anything to be expected in the lifetime of anyone who had graduated from college by 1992. As a consequence, recommendtions about how to proceed must be expected to improve from time to time. What is recommended in 1993 may well be superseded, because we know more, in 1994 or 1995. Our recommendations must have the character of releases of software, where we expect release 3 to be better than release 2 - - and to be eventually supplanted by release 4. For a more detailed discussion, see Tukey 1991p ("Consumer datesware").

What is needed is a consumer product - - something designed by experts for innocents to use.

To most of those who think about better procedures, usually in some version of a single model approach, dirtying one's hands with a consumer product appears to be singularly unattractive - - at least few seem interested in taking part in such a collective effort.

This is most regrettable.

## 5 E is for Engineering

If we think of large parts of the world as diveded into science, engineering (if you prefer, technology) and manufacturing (if you prefer, production), then, in medicine, diagnosis and prescription are engineering, as, in agriculture, is the work of the county agent.

Our concern in this document - - finding, assessing response to challenge of and choosing among procedures - - is engineering also. (Using the procedures is production.)

This may be one reason for disenchantment with consumer datesware. To the extent it is, it is a reason that will not go away.

Given that our concern is with engineering, it follows that we are likely to be concernd with additional attributes beyond those which arise naturally in a science-flavored approach to the underlying problems. The history of the milk bottle is an illuminating example. Maximum milk for minimum glass would lead to spherical shape; adding a bound on cross-sectional dimensions would lead to circular cylinders with hemispherical ends. The following steps of practicality occured, historically:

- a flat bottom to rest stably,
- a reduced top to allow a small cap,
- a rounded square cross section to pack better in the refrigerator,
- pushed in sides to make the bottles easier to grasp,
- replacement of glass by paper restarted evolution with new considerations.

Those who had been concerned with the initial optimization could logically (but not reasonably) have argued that all these additional considerations had dirtied up a clean problem. As a user of milk bottles, I am glad to see the additional attributes brought in.

## 6 F is for *no* Free Lunch

The principle we need to adopt - - and to cling to until there is evidence to the contrary is:

"There is no free lunch!".

# 7 I is for Impalpables

It should be clear that things we cannot touch must fail to help us. In the simple case of a subset of models or challenges offering alternative distributions - - where small sets have small probabilities - - we will learn nothing from, as a result of inference, from the content (theoretical or observed) of a set so small as to contain only a fraction of an observation, which has to mean a combination of (a) a model expectation less than one, and (b) an observed count of zero - - or perhaps one.

A fortiori, passing to the limit, we can make no use in practice, of probability densities as the limits of the content of such small sets. Densities may play some role in identifying interesting procedure (possibly as candidates, more likely as reference points) and as providing candidate behavior for simple procedures faced by simple challenges.

Davies expresses important aspects of the practical uselessness (page 28) as "Statisticians have developed [concepts] such as efficiency, likelihood, sufficiency, Kullback-Leibler discrepancy, Fisher information and admissibility. [These concepts] are *pathologically discontinuous......*silliness is only avoided by an appeal to be reasonable."

It is easy to take things too far. I concur that "likelihodd" and "sufficiency" do not belong in either elementary or intermediate courses, and are certainly unsafe unless applied to individual members of a suite of models that we are considering. These concepts do, however, have useful roles in going from some of the simpler models to interesting (rather than acceptable or trustworthy) alternative candidate procedures. In fact, a "no free lunch" attitude is quite compatible with rather strong reliance, as one criterion of procedure performance, on

$$\text{polyefficiency} = \min\{\text{idioefficiency} \mid \text{models in the suite}\}$$

where "idioefficiency" refers to efficiency assessed within one particular challenge (probably also within constraints like equivariance). These ideas have a place at a sufficiently advanced stage, provided we have given up the single model for the suite of challenges.

The ideas surrounding "Fisher information" and "scores" may eventually have something helpful to say about blandness.

* smoothed densities *

Not only can densities and likelihoods not serve us in the actual inference - - although they may help us in suggesting candidate procedures or in assessing behavior for the simpler procedure-model combinations - - but smooth densities, though plausibly helpful once smoothed enough (smoothed at a scale related to sample size) cannot be helpful unless smoothed enough to involve enough observations for each smoothed value.

The right motto here is not just *de minimis non curat lex* but ex *impalpabilis nihil!*

## 8 J is for Jackknifery

All so-called "resampling methods" are only justified asymptotically - - and thus suffer, to a still uncertain degree, from all the ills of asymptotic procedures.

Like all other asymptotic procedures, they are only used in finite-sample situations, and choices among them need to be based an finite-sample behavior.

The jackknife algorithm can often be perceived as projection of a curved manifold on a flat tangent manifold. As such:

1. it makes no contribution to robustness, and
2. it does nothing to allow for the consequences of skewness.

   The naive bootstrap, in which the empirical percentage points, $y_+$ and $y_-$ of the bootstrap distribution are used as end-points of a confidence interval, makes the effects of skewness twice as bad as they usually are.

3. We do not know enough about the transposed bootstrap, which uses $(2y^{'} - y_*, 2y^{'} - y_-)$ as its confidence interval.

   We also need to be aware that

4. we do not know enough about estimating the impact of skewness on appropriate confidence intervals, (but see Susan Arthur's Thesis).

## 9 L is for Low-probability events

Davies's emphasis on approximation (cp. Section 2 above) is one instance of "de minimis non curat lex" (roughly translatable as "the law pays no attention to things that are too small"). Besides small differences in result we would like to pay no attention to too small probabilities.

We usually need, explicitly or implicitly, to keep such ideas in mind in discussing which quantitative aspects of a model - - or better of a suite of alternative models - -

we should choose as our targets. (This is more difficult in the insurance industry, where relatively catastrophic events of very low probability - - like a major hurricane track across a built-up area of Florida - - have an importance that often outweighs their relatively small probability.) In measurement, however, this phenomenon is usually absent. A sufficiently wild value, for example, simply does not deserve the degree of credence given it by taking the arithmetic mean of the possible values, weighted by their distribution, as the target with we should be concerned.

Another way to put it is that, if we take small probabilities as *de minimis*, the arithmetic mean is discontinuous near $\pm\infty$.

We probably need, therefore, in dealing with distributions, to follow the broad pattern illustrated by:

1. routinely talk, as Bourhaki would say "par abus de language" about "arithmetic means",
2. keep hold of an almost universally applicable caveat that "arithmetic mean" really means "arithmetic mean modified for large values to avoid trouble" (where the modification may, for instance, involve truncation, censorship, or Winsorization),
3. be unsurprised that such realism about targets to be studied appears to have some repercussions on the procedures with which it is reasonable to study these targets.

Downplaying extreme values, then, is likely to be required for both the model or the challenge, and for the data. It is important to recognize that these two kinds of downplaying are in no sense the same, that either would be required in the absence of the other, and that doing either one alone may not suffice.

"Dehorning" *data* values is a reflection of realism about how data is generated - - a recognition that varied mechanisms can cause small numbers of odd values, often best regarded as irrelevant.

"Dehorning" the definitions of targets is a reflection of realism about what we would really do if we really knew the finest details of the "truth" - - of what optimists think models are trying to tell us.

# 10 M is for Multiplicity

Much attention has been given to techniques for dealing with multiplicity. Very little attention has been given to rationales of how multiplicity should be handled and less to why these rationales should arise.

While the interaction between better philosophy for multiple comparisons and the other issue we are discussing here is probably small, it may not be negligible.

As examples of procedural issues relation to multiple-comparison philosophy we can note: (a) Benjamini and Hochberg's approach to the definition of error rates, and (b) the plausibility of some flexibly guided procedures to help judge which (non-significant) appearances deserve mention as "leanings" or "hints".

The thinking about multiple comparisons that we have done has to have substantial impact on the choice of targets for inference. Basically uninformative targets, like "there is demonstrably some difference among the responses to the different treatment" are useless. Thus we are led to choose targets which do have specific consequences, like the maximum deviation of any observed simple comparison from its long-term value.

## 11 P is for Partial models and Partial challenges

* the measurement case *

In most measurement-like problems it is not difficult to separate relevant models into two parts:

- a functional model which describes levels of individual quantities and the ways in which related quantities depend upon one another, and
- a stochastic model which describes, ordinarily in probabilistic terms, how the observed values differ from the simple behavior described by the functional model.

On the one hand, the functional model is often relatively easy to validate, while the details of stochastic model - - which usually involve impalpables - - can be extremly difficult - - really impossible - - to validate. Both Hooke's Law and the simplest sorts of deviations from it are relative easy to validate, while a detailed probabilistic model of how individual observations of tension and stretch relate to long-run typical quantities is often impossible.

In a measurement situation, where the stochastic model describes the difference beetween the observations and the underlying regularity, we ought to handle the two parts of the model quite differently:

- striving to get as good a functional model as we reasonably can, but
- trying to avoid any need for pinning-down the stochastic model in any detail (especially since such pinning-down is usually impossible).

In other words, we should treat the functional model as a matter of scientific description, but the stochastic model as a matter of statistical robustness.

We will usually want to do whatever we can to keep the form of the functional model simple. Accordingly, we are likely to investigate re-expression - - of responses, or covariates, or both - - as a way to simplify a functional model with a high quality of fit.

When we deal with a suite of challenges, we will do what we can to allow us to work with one functional model and a suite of stochastic challenges. Often the price of beeing reasonable in doing this is (a) the inclusion of one or more diagnostics (diagnostic values or diagnostic plots) into the functional model, and (b) treating these diagnostics as providing additional targets.

### * the meaningfully stochastic case *

On the same day on which I first saw Davies's manuscript, I also saw Joel H. Levine's paperback *Exceptions are the Rules: An inquiry into the Methods in the Social Sciences* (Westview Press, 1993). Levine treats, very persuasively, a variety of diverse problems in a common data-analytic framework. In his examples, the stochastic element ist not just a matter of measurement, rather the distributions involved are matters of underlying behavior. Natural variation is much higher than measurement variation.

In some cases we can split each challenge into three parts:

- a functional (non stochastic) model (to be well fitted),
- a stochastic model of natural variation (to be moderately well fitted),
- a suite of stochastic challenges (against each of which we require relatively good performance).

Here our characterization of an attitude toward the middle piece is based on totally inadequate experience. It is also not clear how often we can expect successful in separating the second and third pieces.

In inferential situations, it may be so difficult to separate measurement variation from natural variation that it will be desirable to work with a single stochastic model rather than with two - - one underlying, the other measurement. The technicalities of inference (if any) that we focus on are likely to be matters of bias in responding to underlying behavior, rather than matters of variance in responding to measurement behavior.

And it will be important to take an Ehrenburgian point of view, seeking out evidence for consistency across similar data sets in diverse circumstance - - initial consistency in large-scale behavior, later, *perhaps*, sonsistency in small-scale deviations from the simplest (or simpler) large-scale descriptions.

It would be very wrong indeed to try to treat the two cases - - measurement and meaningfully stochastic - - in a unified way.

In the meaningfully stochastic case, the semiquantitative nature of the stochastic natural behavior - - and a few quantitative targets, are important elements in the scientific description, along with some number of targets of the functional model. The practical difficulties in the estimation of these important quantities are likely to be based upon natural variation, not upon measurement variation. We want to determine the stochastic model well enough for the stochastic targets we are to focus upon to be well enough defined and helpful enough. (Ideas of efficiency and the like are often, at most, of secondary importance.)

* summary *

Almost all models can be split into functional and stochastic parts. If the stochastic part is only concerned with measurement variation, it should be treated by statistical robustness, where we never plan to approach a precise model. If the stochastic part is mainly concerned with natural (rather than measurement) variation, we should treat the corresponding model chioces, just as we treat the model choices relating to the functional part, seeking to do the best the data can support, both alone and after borrowing strength from other data sets.

## 12 R is for Robustness

Classical robustness ought to parallel classical unique-model inference - frequentist or Bayesian - - and thus to be led to *the* optimum procedure, prescribed now in a less definite way. Since people have only been able to do this asymptotically, classical robustness became an asymptotic exercise. In this context, asymptotic analysis was more illogical than usual, because very large samples indeed would change the framework in which we are to work in at least two important ways:

(a) shift to emphasis on questions of bias rather than on questions of variance, and

(b) restriction of alternative models that could considered relevant in view of the data.

Other reasons can also be adduced. In a robustness context, then, asymptotic theory is a very large $n$ analysis that we hope in moderate $n$ situations though we know it is not applicable for very large $n$, since the problem there will be different. It is quite surprising that asymptotic robustness is as helpful as it is.

Since the Princeton Robustness Study [1] (Andrews, et al 1972e), there has been a parallel stream of finite sample robustness work obviously more closely relevant to the analysis of actual data, which does mainly come in quite finite samples. The work here has been diversified, much of it quite empirical, a little of it quite theoretical, [5] (e.g. Morgenthaler and Tukey 1991c). Such work has been mainly confined to the elementary targets of

location and scale, with some extensions to univariate linear regression [6] (e.g. O'Brien Zambuto, 1991 and Cohen, Dalal and Tukey, 1993). [4]
A few specific technical points seem worth recording:

- The usual asymptotic formulation does account for saddle-shaped behavior by asking for at least uniform quality in a neighborhood. As a result, the blandest model in the neighborhood often has a major effect on the choice of an optimum procedure.
- While the techniques of configural polysampling [5](Morgenthaler and Tukey, 1991c) do allow simultaneous optimization at two (demonstrated) or a few (computationally feasible) models, the results of careful thought indicate that the double-optimization results, though not as misleading as single-optimization results, are best used to unterstand the quality of performance of more empirically chosen procedures, rather than as generating procedures to be recommended for use.
- Discussion of robust multiple regression often errs by failing to distinguish natural variation in the $x$-vectors from measurement variation for them (cp. Section 11 above). As a result some approaches are unduly conservative and wasteful of information.

## 13 S is for Symmetrical models and Symmetrical challenges

Davies comes close to being specific about the role of symmetry (of measurement variation) on page 29, her second demand is a lack of bias for "all symmetric F". Two pages earlier, on Page 27, she says: "It would be adventurous to claim it depends on the data being generated by a random mechanism that is symmetric." There is, of course, no conflict between these points, since the role of symmetry is to identify those models where it is clear that we want to estimate. For symmetric models we can be concerned with both bias and variance, but all definitions of bias for unsymmetric models still seem essentially arbitrary. We probably do better by allowing finite-sample robust procedures to give us a sort of analytic continuation from the symmetric to the unsymmetric.

## 14 V is for Variety

The multi-challenge-guided procedure-oriented approach has long ago given up the idea that models are supposed to reflect truth rather than challenge. If we are to study separable procedures, for example, those for level and for variability-of-estimate-of-level, we have no reason for using the same challenges for different parts of the procedure.

Challenges that are effective in challenging the estimation of level, need not be effective in challenging estimation of variability for an already chosen estimate of level, and vice versa.

Even though the partial procedures are to be used upon the same data sets und their results related to one another, there is no reason why the same challenges should have contributed to the chioce of different partial procedures. Variety in challenge-suites is to be expected.

Notice that our example, in which level must be challenged first, and estimation of variability of estimated level later, illustrates the importance of the distinction between primary and secondary parameters.

A suite of challenges really consists of one or more targets to be estimated and a suite of models that challenge the estimations.

## 15 X is for eXamples, chemical or other

Davies, in an effort to make progress on the vital and neglected issues of how classical inference - - or, better, data-based inference - - ought to be connected with real sets of data, chose to discuss chemical examples, including the amount of some chemical - - later mercury - - in a dust or water sample. (This gives the present writer, who was presumed to be a chemist for four or five years, more than 5 decades ago, a possibly unfair advantage.)

If we take the amount of mercury in a water sample as what has been measured, chemists generally would believe:

(1) The number of mercury atoms in the total sample is a fixed unknown number.

(2) No known method of analysis estimates this number without appreciable bias (meaning, perhaps, never $|$ bias $|\leq|$ *mean* $|$ /1000.

(3) Bias in the chemist's analysis, as opposed to any bias to the estimation procedure, can only be assessed by comparison with the results of a better, but still imperfect, chemical procedure. (Or, perhaps, from the results of analyses on samples prepared to contain a known amount of mercury.)

Bear in mind that chemists have taken over the most diverse - - and most sensitive - - methods of assessing amounts by physical properties, including those that involve no "chemistry" as such - - including nuclear magnetic resonance and neutron activation. As a consequence, "chemical analysis" is today an eclectic composite of all kinds of physical-science and engineering measurement. We should thus not be surprised at the converse.

The most careful measurements made are those made in conjunction with any one of the physical sciences (or any branch of engineering), and they all are believed to behave as just described for the amount of mercury in water.

That which *is* measured is defined by the measurement process used. Labelling it with only the name of what is *intended* to be measured ofen seems convenient - - but is almost always misleading, at least to a degree. Saying that an analysis of measured data "estimates" the quantity named as what wished to measure is almost always misleading whenever we need to be involved in the most delicate detail.

### * models for mercury in water *

Davies's position on models for mercury in water is fairly well delineated by these sentences from her page 28.

"If A is the quantity of mercury in a water sample the obvious candidate is the family of Gaussian distributions. There is no question at all of the statistician knowing that the measurements follow a Gaussian distribution, any reasonable persons degree of belief in the correctness of the model is surely zero. The model is *ad hoc.* Nevertheless the Gaussian model as a model for the observations is in general a good one, it is a reasonable approximation."

If we want an all-purpose distribution for the measurement of small quantities of chemicals, I assert:

(1) we will usually do better by ascribing the Gaussian distribution to the *log* of the measured amount rather than the amount itself,

(2) we will be wise to include in the model some form of truncation of the values observed - - or some other way of allowing from the fact that an experienced chemist will in fact jettison sufficiently discrepant values, and

(3) the log Gaussian model can only be really useful as only one of at least a few alternative backgrounds against which to study alternative procedures for analyzing such data (one has only to look at the examples provided to illustrate statistical analysis in some of the major reference books of chemistry to realize that wholly unGaussian outliers do, in fact, occur).

When we are still more realistic, and consider actually a suite of challenges for measurements of small concentrations, we will have to deal with discontinuity near $\pm\infty$ for each and every one of our models (unless, as often happens, the diversity of the models enforces negligible attention to large deviations).

# 16 Z is for fuZZy

A worker trained in terms of single-idiomodel-of-the-truth will have naturally become more unhappy as this account of issues progresses, for everything has been getting fuzzier and fuzzier. As Davies rightly emphasizes, the single idiomodel approach, as a matter of mathematical result rather than as a matter of cultural rigidity - - or possibly of learning from expierience - - was very fuzzy indeed. But the fuzz was routinely swept under the rug - - no one asked all the parallel models, and went along with the Gaussian model as a matter of tradition.

If we are willing to give up the *hubris* of assuming that one can divide the world cleanly and sharply into black an white - - no shades of gray (it seems impossible to find a rational support for this *hubris*), fuzziness in moderation seems a good thing, not a bad one.

We need to work hard to reduce fuzziness where we can, since alternative procedures can clearly be bad choices if we know enough about the behavior of challange-procedure pairs; chioce among others may not be clear, and it is reasonable to allow different analysts to have different chioces.

Significance and negligibility need not be the only two alternatives for an observed difference.

Which challenges need to be taken seriously in analyzing a particular data set may also be a matter of judgment.

These sorts of fuzz are very real - - and intrinsic in the fabric of analyzing data. We must not downgrade approaches that admit to such fuzz. We must instead, downgrade those that do not.

We are NOT pump up our *hubris* and:

- expect data analysis pure and simple to settle the questions of the real world, or
- expect any black-or-white procedure to usefully replace the gray tones of actual data.

# 17 Summary

in brief:

A) Approximation, rather than equality, must be our basis of comparison.

B) Blandness, which implies no attempt to steal a free lunch is an essential characteristic of models useful in studying procedure behavior.

C) Alternative challenges, each of which we must fear, are essential in choosing final procedures.

D) Both theorists and practitioners of procedure choice have obligations to join together to establish current (and therefore evolving) best practice in consumer datesware.

E) The choice of procedures is an engeneering matter.

F) There is no free lunch.

I) In the procedures we actually use we cannot use what we cannot touch (e.g.probability densities) though impalpables may help in the choice process.

J) Jackknifery is asymptotic, along with all resampling.

L) Targets like the arithmetic mean place too much importance on small probabilities of large deviations. It may well be reasonable to talk about "the arithmetic mean" if we always remember that what is meant is "the arithmetic mean suitably modified".

M) Questions of multiplicity may interact weakly with most other issues discussed here, but do have strong influence on choices of targets.

P) Separation of models, or challenges, into a functional part and one or two stochastic parts, can be important. Measurement stochastic shoult be handled as a matter of statistical robustness; functional specification should be done as precisely as possible.

R) Robustness is intrinsic in any multiple-challenge approach, hence in any realistic approach.

S) Symmetry of distribution is useful, not because it is truthful, but because it tells us what target we should estimate (to answer the question of level or location).

V) A suite of challenges really consists of one or more targets to be estimated and a suite of models that challenge that estimation. Different targets for the same data set may deserve different suites of challenges.

X) Davies's examples deserve modification, both as to the labelling of what is to be estimated, and as to the relevant challenges.

Z) Almost all the issues above contribute to the fuzziness/flexibility of the results based upon a specific data set. This is good, not bad.

## References

[1] Andrews, D. F, Bickel, P. J., Hampel, F. R, Rogers, W. H, and Tukey, J. W. (1972e). *Robust estimates of location: Survey and advances.* Princeton University Press, Princeton, NJ.

[2] Arthur, S. P. (1979). Skew/stretched distributions and the $t$-statistic. Ph. D. Thesis, Department of Statistics, Princeton University, Princeton, NJ.

[3] Benjamini, Y., and Hochberg, Y. (1992). A synthesis of new approaches to multiple comparison problems. Department of Statistics, Tel Aviv University, Tel Aviv, Israel. Unpublished.

[4] Cohen, M., Dalal, S. R., and Tukey, J. W.(1993). Robust, smoothly heterogeneous variance regression. *Journ. Roy. Statist. Soc., Series C*, 42(2):339–353, 1993.

[5] Morgenthaler, S. and Tukey, J. W. (eds.) (1991c). *Configural Polysampling: a Route to Practical Robustness.* Wiley, New York.

[6] Zambuto O'Brien, F. L. (1991). Regression. *Configural polysampling: A route to practical robustness.* Chap 10, 133-156.

[7] Tukey, J. W. (1991p). Consumer datesware.*directions in robust statistics and diagnostics,* part 2. IMA Volumes in Mathematics and its Applications, 34. (Werner Stahel and Sanford Weisberg, eds.) 297-308. Springer-Verlag, New York.

NOTE: Letters used with years on John Tukey's publications correspond to bibliographies in all volumes of his collected papers.