SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER 19 16669.4-M	2. GOVT ACCESSION NO. AD-A092646
3. RECIPIENT'S CATALOG NUMBER N/A	
4. TITLE (and Subtitle) 6 Styles of Data Analysis, and Their Implications for Statistical Computing	5. TYPE OF REPORT & PERIOD COVERED REPRINT
	6. PERFORMING ORG. REPORT NUMBER N/A
7. AUTHOR(s) 10 J. W. Tukey	8. CONTRACT OR GRANT NUMBER(s) 15 DAAG29-79-C-0205
9. PERFORMING ORGANIZATION NAME AND ADDRESS Princeton University Princeton, NJ 08544	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS N/A
11. CONTROLLING OFFICE NAME AND ADDRESS US Army Research Office PO Box 12211 Research Triangle Park, NC 27709	12. REPORT DATE 1980
	13. NUMBER OF PAGES 11
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) 12 12	15. SECURITY CLASS. (of this report) Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Submitted for announcement only

DTIC ELECTE DEC 8 1980 S D B

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

AD A092646

DDC FILE COPY

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

288450

80 12 01 115

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ARO 16669.4-M

Styles of data analysis, and their implications for statistical computing

Tukey, J.W., Princeton, USA

Session A1/second paper

Summary

Statistical computing almost inevitably implies special programs, systems, or languages. We are gradually learning how to describe -- and attain -- good practice from such points of view as easy use, input compatibility with people, decent numerical-analysis performance, and even easy maintainability. We must do more of all of this, as I hope everyone will agree. We must also adapt to the needs of the times. This requires looking at the latest styles of data analysis and trying to understand their structure from the *user's* point of view: Not just exploratory and confirmatory, but the pieces these can share and the pieces that must be different. Robust techniques, not just alone but in parallel. Things the computer has yet to learn to do, as well as those it can already do.

Keywords: data analysis, exploratory, confirmatory, diagnostic, middleput, preoutput, data expansion, autonomic judgment, SDAPs.

Every set of special programs, every system, every language reflects, perhaps implicitly, an understanding of one or more styles of data analysis. This is unavoidable. This makes the user happy when the style he wants to use is among those reflected. With relatively few exceptions -- as we must regretfully expect -- today's tools -- programs, systems, languages -- reflect yesterday's styles. It is high time for a fashion show, for an introduction to the styles of the new season.

Robust Techniques.

Some of you may think that robust techniques of analysis is the only major new style. We will see shortly that this need not be so. It is, of course, a very important class of innovations. Here we shall discuss it only briefly and generally, emphasizing that

- for the present at least, we expect to provide the results of *both* a classical *and* a robust/resistant analysis.
- iterative calculations can be expected to occur, perhaps in multiple loops, inside (almost) every robust/resistant analysis.
- we badly need procedures that *find* -- and *report to the user* -- multiple answers.

*Prepared in part in connection with research at Princeton University supported by the U.S. Army Research Office (Durham).

Not one of these three makes the planning of tools easier, but all three have to be faced.

Wants of users.

Users would like a single answer, without the need to think about it. If we satisfy this desire, our particular users will function poorly, and our programs and systems will slowly but steadily get a (well-deserved) bad reputation.

Learning how to convey alternative answers, caveats and warnings in such a way -- very specifically including *in such a format* -- as to combine

- reduced user discomfort, with
- increased user response

is one of the main tasks confronting statistical computing. We have tackled the human interface at input -- at least to a degree -- it is now high time for us to tackle the more difficult human interface at output. (If doing this well requires the techniques we ordinarily relegate to "advertising people", such as motivation research, then we will have to do what is required.)

Data analysis.

Quite the opposite of data reduction, data analysis is pretty well characterized by "making more numbers out of fewer". (Once we say this seriously, the reasonability -- even inevitability -- of parallel analyses of a single set of data becomes clear, since uniqueness is not a natural consequence of "fewer → more".) We only complete reducing to fewer numbers when we have calculated a body of numbers *(part* of our analysis) of which we are willing to say, "we have looked, and found no indication of any further informative structure".

When dealing with a single batch of numbers, for example, we can report only a location number and a scale number IF and ONLY IF we have calculated the residuals and carefully examined them for any informative structure. This means looking, at least

- at the large-scale structure of their distribution -- should there be warnings of stretched (or squeezed-in) tails, of skewness, of bi- or multi- modality?
- at their granularity -- are the values actually reported coarse-grained enough for this to deserve notice?
- (if the values occurred or were observed over time, or doing some other linear variable), is there evidence of any substantial time dependence?

  [probably a few more].

Avoiding the pitfalls of "data reduction" stresses our programs, our computers, and our thoughts, yet it is one of the most important things for us to do better and better.

Basic styles.

For many statistical data analysts and in an increasing collection of areas of application, the distinction between

- exploratory data analysis, AND
- critical or confirmatory data analysis

is quite clear and a part of routine thought processes. For others, this may not be so.

Exploratory data analysis is detective work -- numerical, counting, or graphical detective work -- analysis devoted to finding indications -- the "clues" of data analysis -- of what *appears* to be going on, of what *might* be going on. The detective in a classical detective story is effective when he or she finds many clues, of which some are misleading. A set of exploratory data analysis tools are good, are useful, when they find many indications, not all of which we can be sure about, not all of which will be confirmed when-and-IF we can examine additional data.

Critical data analysis involves the assessment of part of the uncertainty of such indications -- of that part corresponding to the differences revealed in the data that was analyzed. Standard errors, tests of directionality (and occasionally, I fear, even tests of significance that do not involve directionality), and confidence statements all use revealed differences to assess that part of the uncertainty that is calculable from the data. ALL also require good judgment in assessing that part of the uncertainty not likely to be revealed, at least by data limited in those ways in which the actual data is limited.

Much data is inevitably submitted to first exploratory -- whether formal or informal -- and then critical techniques. (Who can analyze the economics of this century free of the exploratory result that there seemed to be a depression in 1929?) We are all aware that such overlap has its problems; we need to recognize that we cannot always eliminate them.

Confirmatory data analysis, as we shall use the term, is critical data analysis on an unexplored body of data believed to be either

- parallel to some body (or bodies) whose exploratory analysis (formal or informal) has suggested an analysis -- and, ordinarily, a focus on certain constants produced in that analysis -- for the data at hand, OR
- of such a form and character that either theory (in a scientific or technological field) or purpose (as often in business or government) prescribes the analysis, OR
- of such a form that some standard (really default) analysis is almost inevitable, OR
- gathered in a carefully planned way with this specific analysis in mind.

The distinction between confirmatory and merely critical analyses is crucial, for the understanding and practice of data analysis. However, since its penetration into statistical computing seems likely to be confined to questions of caveats and automatic warnings, we will not try to discuss it more deeply here.

The tasks of inventors and realizers of statistical computing tools are chiefly directed to processes -- rather than to ambient philosophy. So let us to our processes.

Processes of EDA.

The most helpful, and most important, subdivision of processes of exploratory data analysis divides them into

- *autonomic* data analysis processes -- ADAPs -- that convert data to analyses, AND
- *diagnostic* data analysis processes -- DDAPs -- that look at (aspects of) the results of analysis and endeavor to communicate with the analyst about what can be "seen".

It will often be WRONG to separate ADAPs and DDAPs in the *functioning* of statistical computing tools; it will often be essential to separate them in *thinking* about what is to be done.

* further subdivision *

As we will shortly illustrate, ADAPs themselves usually divide into two parts:

- autonomic data *expansion* processes -- ADEs -- that convert our data into *more* numbers (it will be these that our diagnostic processes are likely to need to feed upon), AND
- optimistic concentrators -- OCONs -- that convert the more numbers into the few that we might be satisfied with if our DDAPs have found nothing further relevant.

Two reasons why this distinction is important are (1) that we may properly choose to pair one of several OCONs with a particular ADE, and (2) it may be wise to have an ADE produce, either actually or potentially, more different things than will be used in any one situation.

* a simple example, ADEs *

If we start with just a batch of numbers, our ADE can reasonably make a variety of typical values (median, midmean, biweight-6, and even mean) and a variety of measures of spread(s, median deviation, s_{fh}, pseudovariances, etc.) and a variety of measures of general distribution (e.g. letter values, which are order-statistic related [Tukey, 1977]). It can also reasonably make one or more kinds of residuals, and may not want to destroy the individual values. This is clearly data *expansion*. We intend such an ADE to make all the standard things that either OCONs or DDAPs might require. (In special circumstances, ADEs with even more diversified outputs may be appropriate.)

* a simple example, OCONs *

OCONs that might well be paired with this ADE might produce, alternatively,

1) a mean and a sample standard deviation,
2) a mean and its standard error,
3) all three of the above,
4) a five-, seven- (or more) number summary (Tukey, 1977),
5) a suspended rootogram, either explicit (Kurtz et al 1965, Tukey 1970-71) or implicit (Tukey 1977, Chapter 17).

* a simple example, DDAPs *

DDAPs that we might want to connect to this same ADE, might include, alternatively, or together:

1) a plebian probability plot or some up-to-date improvement,
2) summarized information, as by g and h (Tukey, unpublished, Hoaglin and Peters, 1979), about distribution shape,
3) ordered values of leaps (differences in adjacent order-statistics divided by differences of corresponding theoretical order-statistic typical values),
4) (if the data was collected in order according to time, space, etc.) plots of residuals in order of collection, both raw and smoothed.
5) and so forth.

Why are these things being produced? As a guide to judgment, as a basis for choice. What choice? The choice of what to do next, of whether or not to output the precutput of the OCON, of what ADEs, OCONs, and DDAPs to apply in the next cycle of exploration (special case: the choice to have no next cycle).

* the choice process *

Today our choices are mainly matters of human choice. Tomorrow there can be large elements of autonomy in our choices. We have to think through our DDAPs with both human and autonomic users in mind. Human choice will often be best fed by displays -- pictures are supposed to be worth many words, often they are worth even more numbers. Autonomic choices may have to be fed by summaries of what would have been displays. For the nearer future, then, autonomic choices are likely to need to be paralleled by human lookings, looking most particularly at whatever aspects of the display summarized for the autonomic chooser are *not* covered by the summaries.

Processes of CDA.

We will do well to think of our processes of critical/confirmatory data analysis as following after a sequence of EDA cycles. Indeed we can usually think of hitching a

CDAP -- a critical data analysis process

to an ADE-OCON pair as the typical way to do CDA. Where reasonable, we will want to use a general CDAP.

There are now a number of kinds of general CDA approaches, including:

- differences from piece to piece, implemented with Student's *t*, Wilcoxon or even biweight procedures,
- jackknife procedures, usually with Student's t,
- half-sample procedures (paired or not)
- bootstrap procedures.

Tomorrow there well may be more.

* further example, OCON-CDAP *

If we are to hitch our chosen CDAP onto an OCON, the output of our OCON

must be extensive enough. If the data are so structured to make a factorial analysis of variance reasonable, it will NOT, for example, suffice for the OCON to provide only the analysis of variance table. So many have so often criticized papers that do not give the estimated effects for the various treatments. OCONs that fail in this way fail miserably.

Indeed, we may well want our OCON to carry out the aggregations and poolings discussed and illustrated in Green and Tukey (1960). It would then report the condensed anova table and the effects and interactions that remain apparently relevant.

Process and reality.

We have been describing the logical steps of a data analysis. A statistical computing system need not operate in the way these steps would naively suggest. Steps may only be carried out when their results are needed. Results to be used twice or more may be freely stored or equally freely forgotten and recomputed.

Understanding how to structure the calculations and rememberings may appropriately be a quite separate process, but it will fail to give us the support we need unless it is thought through in terms of a logical and relevant understanding of the steps of the data analysis, some of which we have just described.

We dare not constrain implementation; we must constrain attitude and understanding.

Some verbal schematics.

With this caveat that we are NOT trying to describe implementation, we can go ahead with some schematic descriptions. As we do this, we will find it helpful to have words for at least three kinds of intermediate results:

- *preoutput*, describing what may later be used as either *output* or input to another step (mainly, here, from OCONs)
- *middleput*, describing extensive material intended for another step (mainly, here, from ADEs and to DDAPs)
- *diagnostics*, describing material to be offered to guide choice, either human or autonomic.

The uses of these are different enough that it seems likely that they will be implemented differently.

The basic schematic for an ADAP -- an autonomic data analysis process is, then

input → ADE → middleput → OCON → preoutput

To the arrow coming down from the middleput we would usually attach either

autonomic choice ←
human scrutiny ← } ← DDAP

or, temporarily,

human choice ← *DDAP*

Here choice would set up the next step, specifying ADEs, OCONs, DDAPs (and possibly Autonomic Judgers) as well as making decisions about which preoutputs, already (logically, but not necessarily in implementation) generated, are now certain to be output.

Notice the plurals "ADEs, OCONs, DDAPs". Any one step may involve more than one of each kind. Several in the same step may represent *either* deep understanding of what is needed *or* scratching around in the dark.

As we get more used to alternative outputs and alternative ADAPs, we will find ourselves more and more in need of

SDAPs -- selective data analysis procedures

in which the results of 2 or more (usually more) approaches are examined autonomically, with the result that some (maybe more, maybe all) of these results are passed on or outputted. Here we have almost no experience, so Colin Mallows and I are trying to produce a good SDAP for the problem:

data structure = a batch

objective = shape of distribution .

Time will tell.

Some pictorial schematics.

We close this discussion with some pictures of the flow of information and control in

1) a single ADAP
2) a step of EDA
3) an extended EDA process

for which see exhibits 1, 2, and 3. Remember that the elements of these schematics are logical steps and need not reflect specific implementations or specific choices of time at which things are calculated.

Size of interaction.

At least until autonomic judgment is developed far beyond its present level, the discussion above assumes human intervention at suitable intervals, neither too close together nor too widely separated. I consider heretical both:

- the idea that an analyst should specify each step of his/her analysis, one after another -- this assumes that planning parts of analyses is much easier than in truth it is; that every user will, for instance, instinctively do the right numerical analysis.
- the idea that a package should take the data away and come back with the answers

-- this assumes that planning an overall analysis is much easier than in truth it is.

The proper spacing between human interventions will slowly grow as the years and decades pass by, but, whatever the epoch, it will always be possible *both* to intervene too frequently *and* to intervene too infrequently.

Keeping intervention-spacing roughly tuned to our insights and capabilities will be a challenging important problem throughout the foreseeable future.

Multiple answers

We stressed the need for multiple answers in connection with robust-resistant methods. This need existed when only classical procedures were being thought of; it will exist in the far future, when, perchance, all the procedures we now know have been replaced.

We need only look at multiple regression without prespecification of which carriers (out of a specified collection) are to be used. The methods of Furnival and Wilson [1974] make it quite feasible to learn both which subsets appear to do best and how well they appear to do. (If we have only 10 carriers, say, the methods of Daniel and Wood [1971(1980)] will allow us to look at all $2^{10} = 1024$ possibilities.) Why were users and analysts so willing to *demand* multiple answers here?

I suggest that the same reasons will apply to wider and wider areas of analysis as we come to recognize the nature and diversity of the possible analyses of each of many kinds of problems. Consider multiple regression on a specified set of carriers as an example. The development of techniques for identifying "high-leverage points" has now been extended (Andrews and Pregibon 1978) to the identification of "high-leverage groups", and will inevitably extend to procedures for clustering (plausibly on x's and y's together) all the points in high-leverage entities. If there are k such clusters (some or all may be single points) there are 2^k regressions, one obtained by setting aside each subcollection of these clusters. I suspect that procedures for:

- telling us about all 2^k regressions, including their apparent behavior at each cluster,
- sorting out, algorithmically, those of the 2^k which seem intrinsically most likely to interest us, AND even for
- blending together regressions for different subcollections that lead to seemingly -- but far from certainly -- different regressions

will, in due course, prove to be as useful here as results for multiple subsets have proved to be in the carriers unspecified case.

In a word or two, I believe that "points unspecified" makes as much sense as "carriers unspecified" and that both will always be needed. (At least until they are subsumed into still more flexible descriptions of what is to be done.)

Actual implementation

The descriptions above have been wholly human-directed, emphasizing input, choices, and output. (As I am not a system designer, it would be silly if they were not.) We have valued that they were not intended to describe implementation, but some examples to emphasize this are not likely to be out of place.

We have described our OCONs as chosen at the same times as our ADEs. This does not imply that they need be implemented at the same internal time as their ADEs. They only exist to feed either OUTs, CDAPs, or SDAPs. What is required is only this:

- when their preoutputs are called for, they will be returned.

This need not require us to store the preoutputs themselves. Storing any one of:

- their preoutput
- middleput and OCON (implicit or explicit), ready to make preoutputs, OR
- input, ADE, and OCON, ready to make preoutputs

can service the need. Which to do is a system designer's choice.

That the user cannot tell directly which of these has been done is a proper demand, levied by the user community on the system designer.

Acknowledgements

The author joins with his secretary, Mary Bittrich, in thanking both Brian Kernighan for the use of the experimental program that produced exhibits 1 to 3 and Paul Tukey for guidance through its details.

References

Andrews, D. F. and Pregibon, D. 1978. Finding the outliers that matter. *J. Roy. Statist. Soc. B*, 40:85-93.

Daniel, C. and Wood, F. S. 1971 (2nd ed. 1980). *Fitting Equations to Data*. New York: Wiley.

Furnival, G. M. and Wilson, R. W. 1974. Regressions by leaps and bounds. *Technometrics* 16:499-511.

Green, B. F. and Tukey, J. W. 1960. Complex analysis of variance: general problems. *Psychometrika* 25:127-152.

Hoaglin, D. G. and Peters, S. G. 1979. Software for exploring distribution shape. *Proceedings of Computer Science and Statistics 12th Annual Symposium on the Interface*, J. Gentleman (ed.), Waterloo, Ontario: University of Waterloo.

Kurtz, T. E., Link, R. F., Tukey, J. W. and Wallace, D. L. 1965. The future of processes of data analysis. In *Proc. 10th Conf. Design Expts. in Army Res. Devel Testing*, 691-729.

Tukey, J. W. 1970-1. *Exploratory Data Analysis*, Limited Preliminary Edition (3 vols.) Reading (Mass): Addison-Wesley. (Microfiche available from University Microfilms.)

Tukey, J. W. 1977. *Exploratory Data Analysis*, Reading (Mass): Addison-Wesley.

Exhibit 1

A single ADAP
(3 = input, 4 = middleput, 2 = preoutput,
order of numbers is qualitative order of amount of data)

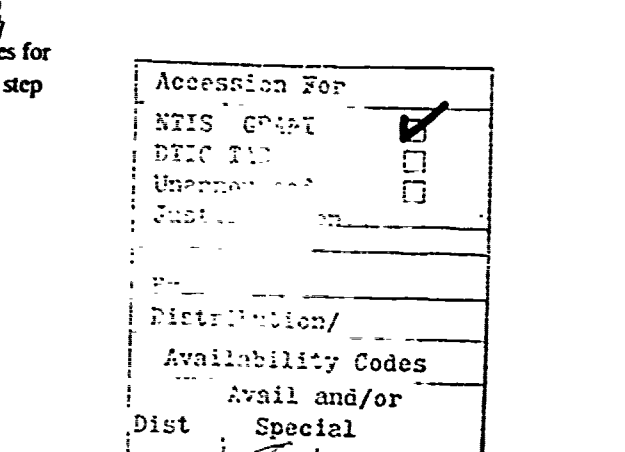

Exhibit 2

A step of EDA
(code as above, also 1 = output and 5 = diagnostics)

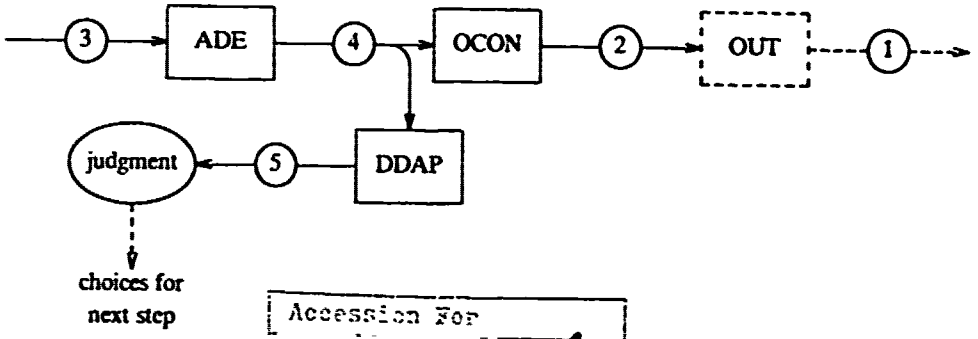

Exhibit 3

An extended EDA process
(dashed lines show implementation of judgment)