Statistics 215a – 9/15/03 – D. R. Brillinger

The landmark paper on EDA:

J. W. Tukey (1962). The future of data analysis. *Ann. Math. Statist.* 33, 1-67.

I. General considerations

For a long time I have thought that I was a statistician, interested in inferences from the particular to the general

Data analysis includes:
i) procedures for analyzing data, techniques for interpreting the results of such procedures,
ii) ways of planning the gathering of data to make its analysis easier, more precise or more accurate
and
iii) all the machinery and results of (mathematical) statistics which apply to analyzing data.

Large parts of data analysis are inferential …, but these are only parts

Statistics has contributed much to data analysis, e.g. asymptotic power, decision functions, properties of non-normal samples

Least squares goes back more than a century and a half, Gauss (1803)

The last century has seen great developments in regression techniques

… the great innovations in statistics have not had correspondingly great effects upon data analysis

… we need to stress flexibility of attack, willingness to iterate, and willingness to study things as they are

We should seek out unfamiliar summaries …

The comparison under suitable or unsuitable assumptions, of different ways of analyzing the same data for the same purpose has been a unifying concept in statistics.

Many seem to find it essential to begin with a probability model containing a parameter, and then to ask for a good estimate of this parameter … Many have forgotten that data analysis can … precede probability ,models, that progress can come from asking what a specified indicator … may reasonably be regarded as estimating.

M. B. Wilk “intuitive generalization”

what makes a science:
a1) intellectual content,
a2) organization into understandable form

a3) reliance upon the test of experience as the ultimate standard of validity (statistical science)

data analysis must:
b1) seek for scope and usefulness rather than security,
b2) be willing to err moderately often in order that inadequate evidence shall more often *suggest* the right answer
b3) use mathematical argument and mathematical results as bases for judgement rather than as bases for proof or stamps of validity

M. B. Wilk, “The hallmark of good science is that it uses models and `theory’ but never believes them.”

Dangers of optimization.

G. Kimball “There is a further difficulty with the finding of `best’ solutions. All too frequently when a `best’ solution to a problem has been found, someone comes along and finds a still better solution simply by pointing out the existence of a hitherto unsuspected variable. … The time is better spent in real research…”

In data analysis we must look to a very heavy emphasis on judgment.
a1) judgment based upon the experience of the particular field of subject matter from which the data come,

a2) judgment based upon a broad experience with how particular techniques of data analysis have worked out in a variety of fields of application,
a3) judgment based upon abstract results about the properties of particular techniques, whether obtained by mathematical proofs or empirical sampling

Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise.”

II. Spotty data – a growth area

errors, blunders, outliers, wild shots, large deviations

rejection of observations

P. Olmstead “engineering data typically involves 10% of ‘wildshots’ or ‘stragglers.”

decreases chances of real effects and increases the number of false negatives

in formalizing stick to symmetric distributions

trimming – removing equal numbers of the lowest and highest observations

winsorizing – replacing the most extreme observations by the nearest unaffected values

Student’s t went into service on the basis of some empirical sampling and some Pearson curves – it filled an aching void

Geary “Normality is a myth; there never has, and never will be, a normal distribution.”

III. Spotty data in more complex situations

a1) if a particular deviation is much “too large”, its effect upon the final estimates is to be made very small

a2) if a particular deviation is only moderately “too large”, its effect is to be decreased, but not made negligible

I look forward to the automation of as many standardizable procedures as possible

FUNOP

FUNOR-FUNOM

IV. Multiple-response data – a growth area

Cochran “regression is the worst taught part of statistics”

We do not have a good collection of ideal, or prototype multivariate problems and solutions

taxonomy, classification (data mining)

V. Some other promising areas

stochastic-process data

b1) decades were lost because over-simple probability models in which there was a hope of estimating everything were introduced and taken seriously

“hidden replication”

For the most useful information is not infrequently found to reside in the apparent wild shots themselves

VI. Flexibility of attack

choices of modes of expressions

Indications are not to be judged as if they were conclusions

Multiple comparison methods

VII. A specific sort of flexibility

the vacuum cleaner (and attachments)

the calculation of residuals

non-additivity

The simple graph has brought more information to the data analyst’s mind than any other device

FILLET

VIII. How shall we proceed?

the purpose of data analysis is to analyse data better

the necessarily approximate nature of useful results in data analysis. Our formal hypotheses and assumptions will never be broad enough to encompass the actual situations.

The need for a free use of ad hoc and informal procedures in seeking indications

We must plan to learn to ask first of the data what it suggests, leaving for later consideration the question of what it establishes

the role of empirical sampling, Monte Carlo

data analysis has … to be an experimental science

Statistics 215a – 9/15/03 – D. R. Brillinger

Linear fitting

OLS. Single explanatory – simple regression

dependence of Y on X

straight line, $Y = \alpha + \beta X$

data (x_i, y_i), $i=1,\ldots,n$

$$\min_{\alpha,\beta} \sum_i (y_i - \alpha - \beta x_i)^2$$

normal equations

fit $a + bx_i$

relationship?

residuals $y_i - a - bx_i$

patterns?, stleaf, dependence?

orthogonalities

data = fit + residuals

lm()

Difficulties of interpretation
   Anscombe's data

OLS. Several explanatories

   X_{ij}, $i=1,\ldots,I$; $j=1,\ldots,J$

WLS.

   $\min_{\alpha,\beta} \sum_i w_i(y_i - \alpha - \beta x_i)^2$

NLS.

$$Y = g(X;\beta)$$

$$\min_{\beta} \sum_i (y_i - g(x_i;\beta))^2$$

fit $y = g(x;b)$

residuals $y_i - g(x_i;b)$

nls()