John Nelder: Breadth and depth in statistics

David J. Hand
Imperial College, London
and
Winton Capital Management

28th March 2015

John Nelder: Master of Synthesis

Simplex method: 1965: *A half century !*

"A method is described for the minimization of a function of n variables, which depends on the comparison of function values at the (n+1) vertices of a general simplex, followed by the replacement of the vertex with the highest value by another point."

Nelder and Mead, Computer Journal, 7, 303

20,290 citations

General balance

On starting at Rothamsted John was given the task of analysing some experiments on trace elements

- Complex with crossing, nesting, confounders, etc

Led him to develop the theory of *generally balanced designs*, enabling the correct form of analysis to be derived from the design

Senn: John's development encompasses:

> "*completely randomised designs, randomised blocks, split plots, Latin and Graeco-Latin squares, split-split plots, balanced incomplete blocks, balanced lattices, Youden squares, and many more, in fact all designs possessing the property of 'first-order balance'*."

Generalised linear models

"The technique of iterative weighted linear regression can be used to obtain maximum likelihood estimates of the parameters with observations distributed according to some exponential family and systematic effects that can be made linear by a suitable transformation"

Nelder and Wedderburn, JRSS-A, 135, 370

Book: 22,523 citations

Hierarchical generalized linear models

GLMs + non-normal random effects

Lee and Nelder (1996, 2001, 2006, ...)

John saw the rise of the Bayesians

- and he saw the impact of computers
  - but he didn't witness the sudden and dramatic interest in "big data"
    - and to a lesser extent "data science"

but ....

Presidential address 1986

"Statistics, science and technology"

Link between statistics on the one hand
and science and technology on the other

"… the ideas of statistics and the work of statisticians are not having the influence that they should on the work of scientists and technologists."

John's perspective

1) Statistics is central to *matching theory to data*
2) Good ***design*** is critical:

   We all have "*horror stories about studies where no useful conclusions could be drawn because of poor or non-existent design*"

"*As we look back over the history of our subject we can see swings towards views of statistics that are either data-heavy or model-heavy*"

1800s: data was central: *aliis exterendum*
1900s: analysis, mathematics, inference dominated
2000s: data makes a comeback

The resurgence of data-centric views

1. Tukey introduced the term "*data analysis*" (c. 1977)

2. John describes a meeting in north America
   "*where a speaker explicitly distinguished between statisticians and those who looked at data*"

→ Perhaps, to a large extent, we have only ourselves to blame for the facts that data mining and big data are largely owned by computer scientists rather than statisticians

John called for more attention to be given to *data*, when training statisticians

He argued ***against*** the notion that this is the sort of thing you pick up on the job ...

... claiming that, on the contrary "*the reverse is the case, and ... it may be easier to pick up some unfamiliar theory later*"

Perhaps John's interest and expertise in ***design*** meant he did not foresee the "big data" revolution

Because the key thing is not the fact that the data are "big", [whatever that means]

But that the data are mostly ***observational***

- that they are a ***side effect*** of some other exercise
- that they are often, for example
  - ***administrative*** data
  - ***transactional*** data
  - ***automatically*** measured

This is of ***fundamental*** importance

- data not designed to answer the question you want to
- selection biases
- quality issues

Statistical education

John did not spend a large chunk of his career in education

But he still had views on it – gained, I suspect, at least partly from his time at Rothamsted appointing statisticians

He criticised statistical writing and education for focusing on "*the analysis of the unique experiment or study*"

And the emphasis on seeking "*differences*" (as in a basic classical two-sample *t*-test) rather than the scientific perspective of seeking significant "*sameness*" – that is reproducible results

John had a view on the role of mathematics in statistics

"The main danger, I believe, in allowing the ethos of mathematics to gain too much influence in statistics is that statisticians will be tempted into types of abstraction that they believe will be thought respectable by mathematicians rather than pursuing ideas of value to statistics"

"... whereas a topologist, say, may justifiably aim his papers at other topologists, and not try to explain himself to number theorists, statisticians need to take a different approach."

The computer

John witnessed the entire statistical revolution consequent on the development of the computer

Described three generations

- isolated programs
- statistical packages (and languages)
- statistical expert systems

Key question: who is the user?

Misuse arising from ease of use

"*regression analysis is being used in foolish ways in the vicinity of almost every computer installation*"

[Robert Hooke, 1980]

Motivating the development of user friendly and protective front ends

John distinguished between *authoritarian* and *libertarian* systems:

- The former constrained the user to a limited set of options at any stage (e.g. a pull down menu system)
- The latter concentrated on giving good advice conditional on the path already taken

GLIMPSE

"A knowledge-based front end to the statistical package GLIM 3.77"

> *"Designed to provide both semantic and syntactic help and advice on the statistical strategy of generalized linear modelling"*
>
> *"must deal with a changing knowledge base due to the discovery of new facts gleaned during the course of data analysis. It must also cater for users with different levels of expertise"*

The scope is illustrated by John's list of ways a user could query his GLIMPSE front end:

i) what options are available at this stage?

ii) what does a word mean in a question?

iii) why is this question being asked?

iv) what does the system advise?

[[John grasped complex ideas with ease, and it was sometimes necessary to go through an extra loop to bring them down to the level of mere mortals. I can recall someone commenting that GLIMPSE itself needed an 'ordinary user' friendly front end]]

However, ...

the potential of statistical expert systems did not materialise, at least in the way expected

However2, ...

A new dawn?

Zoubin Ghahramani's $750,000 Google grant:

"*The ultimate aim of the Automatic Statistician is to produce an artificially intelligent (AI) system for statistics and the data sciences.*"

Fifteen years later:

From statistics to statistical science, 1999

JRSS-D, 48, 257

"*The subject should be renamed **statistical science** and be focused on the experimental cycle, **design-execute-analyse-predict***"

> Again arguing that statistics belongs with science, and not with mathematics

"*I believe that all statisticians should have contact with experiments during their training*"

Drew attention to the iterative nature of real analysis, in contrast to the linear style implied in many statistical texts

Derided the p-value culture

"The kernel of these non-scientific procedures is the obsession with significance tests as the end point of any analysis"

"The most important task before us in developing statistical science is to demolish the P-value culture"

Very much aware of the selection bias issues which have been attracting so much attention recently

What should we do? [And remember, written in 1999]

1. Sort out modes of inference

"At least once a year I hear someone at a meeting say that there are two modes of inference: frequentist and Bayesian. That this sort of nonsense is so regularly propagated shows how much we have to do."

He went on to say that he belongs to the "*flourishing school of* ***likelihood*** *inference*"

2. Remove non-scientific procedures

He's thinking of replacing *p*-values by estimates of effects and their uncertainties

and also refers to discarding multiple comparisons tests and playing down distribution-free methods

3. Promote good software

4. Teach statistical science

"*Embed courses for both statisticians and experimenters in the experimental cycle*"

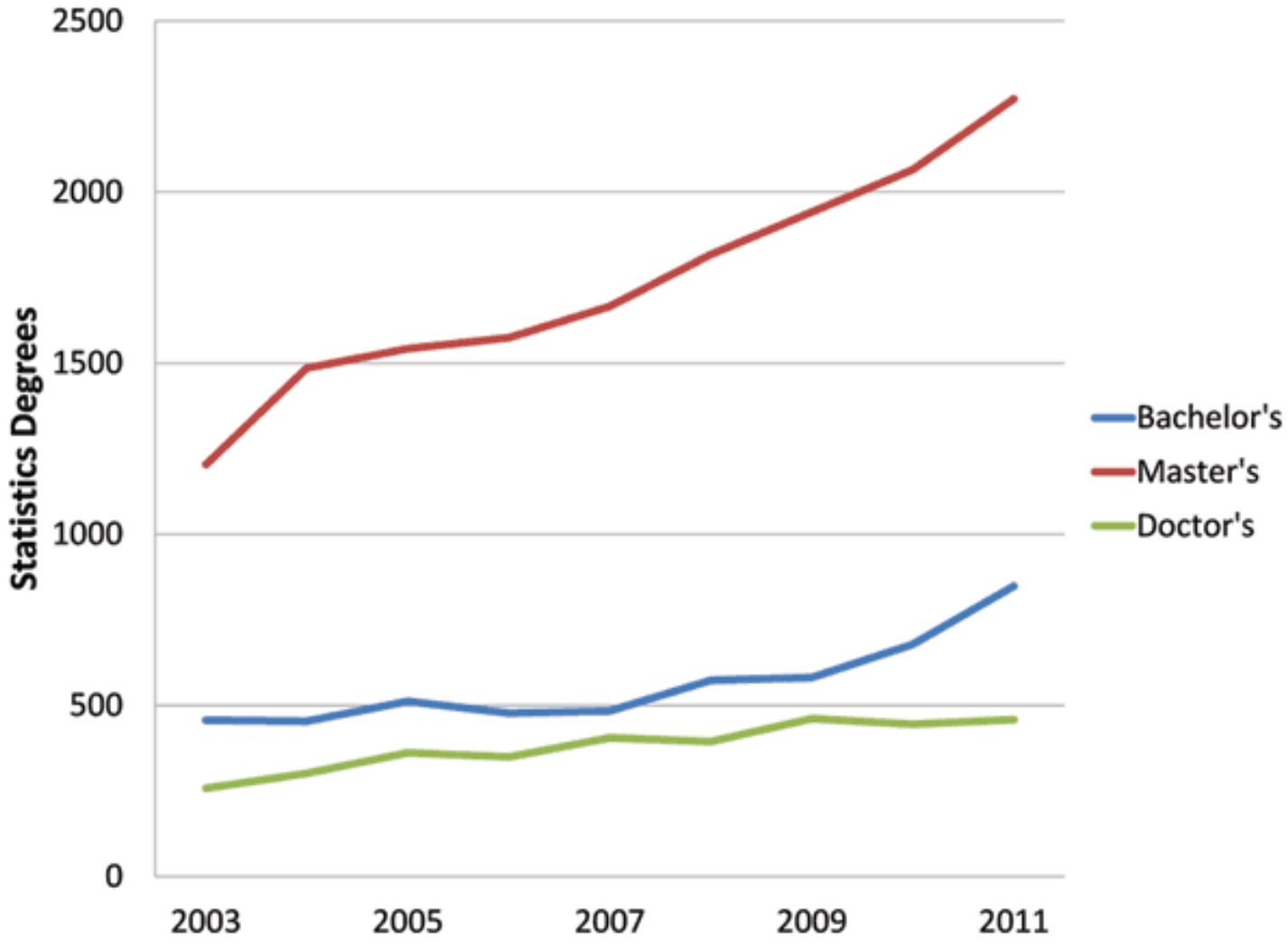
He concludes by quoting Deming:

"***You do not have to do any of these things; survival is not compulsory***"

Particularly relevant in today, in a world of "big data"

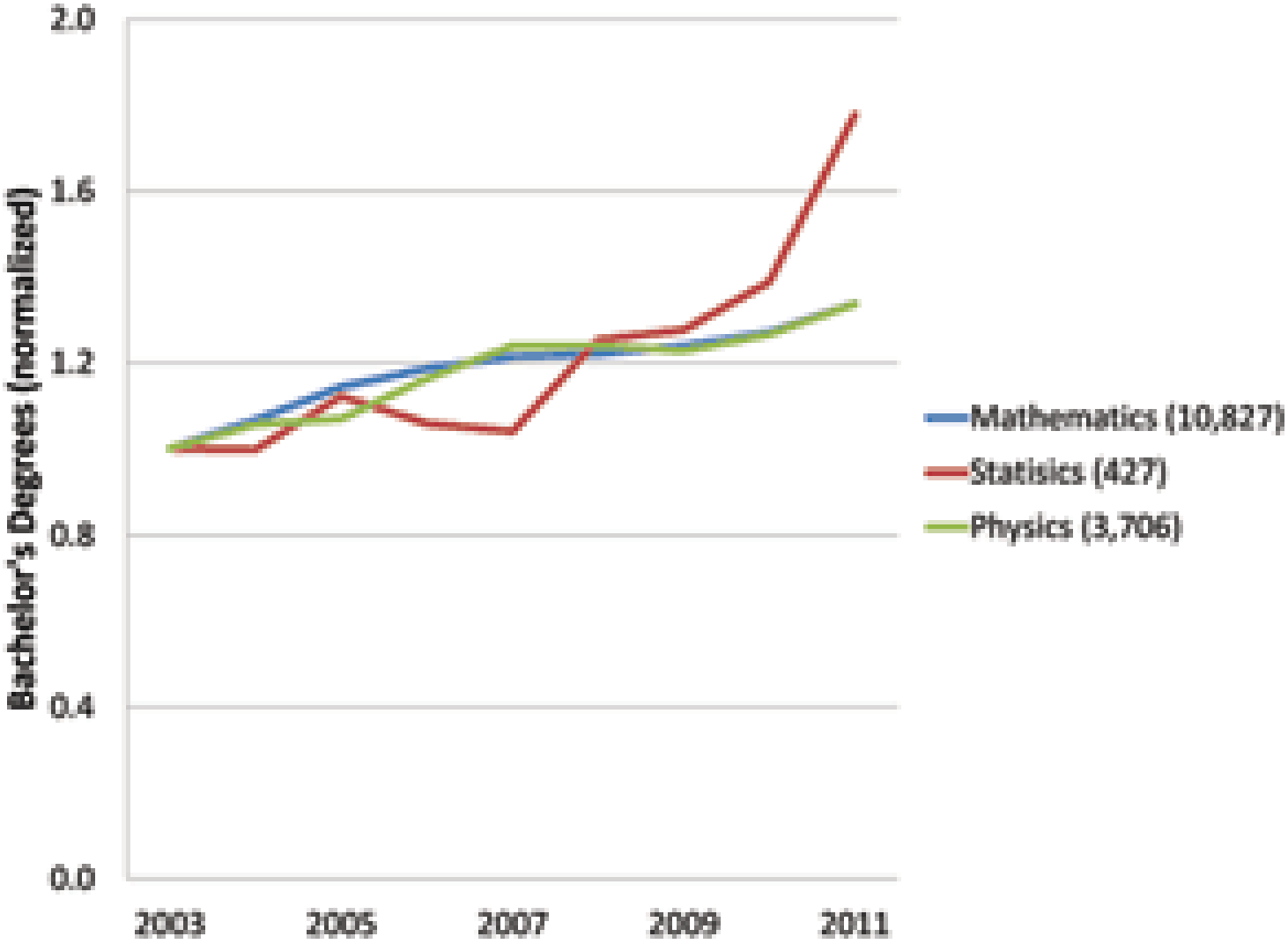
But one thing has changed since 1999:

One of the discussants referred to "the number of unfilled university places [for statistics]"



Statistics degrees in the US

http://magazine.amstat.org/blog/2013/05/01/stats-degrees/



Bachelor's degrees in US for maths, stats, and physics

Conclusion

Breadth and depth, certainly

But above all, ***synthesis***