Methods of Applied Stats, Introduction and data science

Patrick Brown, University of Toronto and St Michael's Hospital

September to December 2020

Outline for today

- Course overview
- Statistics vs Data Analysis
- Examples
- Generalized Linear Models
- Likelihood-based inference
- Applications
- Scientific writing

Web pages

- q.utoronto.ca/courses/182130
- pbrown.ca/teaching/appliedstats/data

What is Applied Statistics

- Mathematics: The abstract science of number, quantity, and space, either as abstract concepts (pure mathematics), or as applied to other disciplines such as physics and engineering (applied mathematics) Oxford Dictionary
- Statistical Sciences: a subset of the Mathematical sciences concerned with uncertainty and randomness. (no citation)
- Applied Statistics: subset of Statistical Sciences, using data to solve problems and answer questions where *uncertainty* and *randomness* are involved. (no citation)

The 'uncertainty and randomness' part is important.

- All Applied Scientists either collect data, manage data, or use data to solve problems.
- Applied Statisticans make *inference* about phenomena which haven't been observed directly.

Data Science

From D. Donoho (2017). "50 Years of Data Science". University of Michigan's Data Science Initiative

> *This coupling of scientific discovery and practice involves the collection, management, processing, analysis, visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of scientific, translational, and interdisciplinary applications.*

- Applied Statistics is a subset of data science
- Not all Data Science is Statistics
  - much of Data Science concerns things which are observed directly
- Statisticians need a good number of non-Statistical data science skills
- Is the distinction between Statistics and Data Science important?

P. J. Diggle (2015). "Statistics: a data science for the 21st century". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178.4, pp. 793–813. DOI: 10.1111/rssa.12132

Statistics: a data science for the 21st century

Peter J. Diggle

Lancaster University, UK

[*The address of the President delivered to* The Royal Statistical Society *on Wednesday, June 24th, 2015*]

Summary. The rise of data science could be seen as a potental threat to the long-term status of the statistics discipline. I first argue that, although there is a threat, there is also a much greater opportunity to re-emphasize the universal relevance of statistical method to the interpretation of data, and I give a short historical outline of the increasingly important links between statistics and information technology. The core of the paper is a summary of several recent research projects, through which I hope to demonstrate that statistics makes an essential, but incomplete, contribution to the emerging field of 'electronic health' research. Finally, I offer personal thoughts on how statistics might best be organized in a research-led university, on what we should teach our students and on some issues broadly related to data science where the Royal Statistical Society can take a lead.

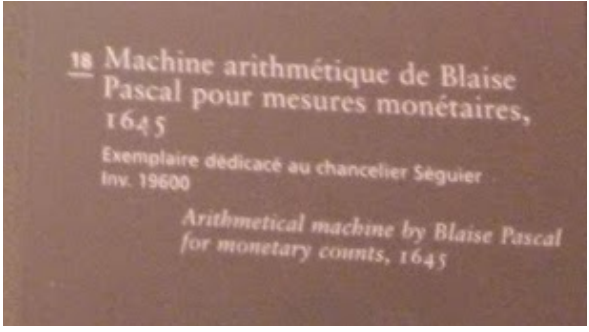
Keywords: Data science; Electronic health research; Health surveillance; Informatics; National Health Service prescribing patterns; Reproducible research; Statistical education

. The rise of data science: threat or opportunity?

'he first thing to say is that we have been here before. I began my career in 1974, at which t
tatistical software packages were beginning to become widely available. This was seen by s
f my colleagues as an existential threat. If useful statistical methods could be implemente
oftware, surely would not the need for statisticians diminish? In fact, the reverse happe
or at least three reasons. Firstly, if something is impossible it is easy to convince yourself
ou can get by without it. Packages enabled scholars of many disciplines who might previo
ave considered statistics irrelevant to their subject to begin to appreciate its power. Secon
ackages enabled *statisticians* to do more things routinely, again increasing the reach of stati

Early Computing

- Blaise Pascal, 1645
- Charles Babbage, Ada Lovelace 19th C
- in the 1940's "you hired a person, usually a woman, who used a calculating machine' '
- "they used women plural, who were called 'computers' and whose collective job, in production line style, was to turn a data set into an analysis of vaiance, each computer having been trained to carry out a specific task.' '
- Colossus machine, Bletchley Park code breaking, February 1944



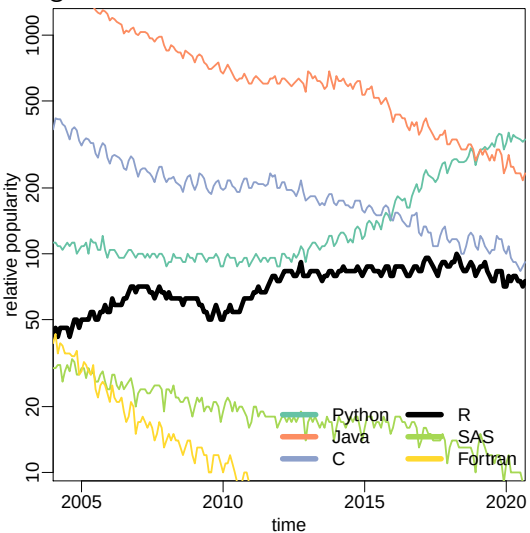
Modern statistical computing

- Late 1960's GenStat and SAS
- "early 1970s, programming was beginning to enter the statistics curriculum''
  - "Nelder and Wedderburn's (1972) breakthrough paper on generalized linear models, and its dissemination through the GenStat and GLIM packages. This development offered, for the first time, a transparent path from the theory of the exponential family, through the unifying framework of the iteratively weighted least squares algorithm to the implementation of a wide range of statistical methods in a single piece of software.''
- "Gelfand and Smith (1990) brought MCMC methods into the statistical mainstream.''
- "early versions of the BUGS software were running from 1991 onwards, before the first stable version was released in 1995.''

R: The Guv'nor!

- "Arguably the most transformational development in statistical software since the 1990s has been the R project (www.r-project.org).''
- "The R language, which had its origins in the S language (Becker et al. 1988)''
- "One important aspect of R is that it is open source''
- "its crucial feature is its extendibility through a plethora of 'contributed packages"''

Google searches



Threats to Statisticians

- Statististical software was seen as a threat in the 1070's
  - "If useful statistical methods could be implemented in software, surely would not the need for statisticians diminish?' '
  - "In fact, the reverse happened' '
- Is Data Science a threat?
  - "Undoubtedly, there is a threat, but it is one that has been with us for a very long time, namely that any numerate scholar can operate as an amateur statistician within their own substantive discipline.' '

What Diggle says

- What can we *[statisticians]* offer *[to data science]*?
- Crucially, we can assert that uncertainty is ubiquitous
- probability is the correct way to deal with uncertainty
- We understand the uncertainty in our data by building stochastic models, and in our conclusions by probabilistic inference.
- we also minimize uncertainty by the application of the design principles that Fisher laid down 80 years ago
- Also, context matters. [...] the extraction of knowledge from a given set of data depends as much on the context in which the data were collected as on the numbers that the data set contains.

Diggle (2015)

Statistical Computing post WinBUGS

INLA

- H. Rue et al. (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations". In: *Journal of the Royal Statistical Society B* 71.2, pp. 319–392. DOI: 10.1111/j.1467-9868.2008.00700.x
- R-INLA package r-inla.org
- fits a range of latent Gaussian models with a nice syntax
  - `y ~ x + f(time, model='ar1')`
- Less labour-intensive than MCMC

Statistical Computing post WinBUGS

Stan

- mc-stan.org, Rstan package
- named for Stanislaw Ulam, Aug 2012
- WinBUGS-style models files
- Hamiltonian MCMC, better than WinBUGS
- Spatial statistics not well supported

Automatic Differentiation

- 'frequentist-style empirical Bayes'
- ADMB, Automatic Differentiation Model Builder, Fournier et al. (2012)
- TMB, Template Model Builder, Kristensen et al. (2016)

Statistical Computing post WinBUGS

Variational Bayes

- Popular amongst computer scientists
- Less accurate than INLA (citation needed)

R interfraces

- wrappers for formula-style models
- bmrs, rstanarm (for stan)
- glmmADMB, glmmTMB

Beyond R?

Python

- 'bigger' than R
- NumPy package for scientific computing
- R's better at Statistics
- Python's better at data analysis

Julia

- development started in 2009, first released 2012
- very fast

Approaches to Applied Statistics

- L. Breiman (Aug. 2001). "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)". In: *Statist. Sci.* 16.3, pp. 199–231. DOI: 10.1214/ss/1009213726
- Data modelling culture vs Algorithmic modelling culture
- Parametric models vs machine learning
- Extracting information vs prediction
- Minimise bias vs minimise variance
- A third culture: experimental design and survey sampling

What is Big Data?

- Error = Bias + Variance
- Bias = error because we're using the wrong model
- Variance = error because there's uncertainty in our parameter estimates
- Variance decreases with N, bias doesn't.
- Conventional data:
  - Bias = $\epsilon \cdot$ Variance, or at least this is implicitly assumed
  - Error $\approx$ Variance
- Big data:
  - Variance ≈ 0 for the 'standard' statistical methods
  - Error $\approx$ Bias
- Note the lack of citations

Methods of Applied Stats

- Applied Statistics rather than Data Science
- Focus on data modelling, rather than design or machine learning
- Accommodate Big Data with Bigger (but still parametric) Models
- Emphasise models for non-Gaussian data
- ...which often requires Bayesian Inference

What's covered in this course?

- Linear mixed effects models
- Bayesian hierarchical models
- semi-parametric models
- case control studies

Some quotes

- 'Solving problems, not analyzing data' Diggle
- 'All models are wrong, some are useful', Box
- 'We buy information with assumptions' Coombs
- 'An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.' Tukey
- 'A goodness-of-fit test is a measure of how much data you have.' Diggle
- 'If your experiment needs statistics, you ought to have done a better experiment.' attributed to Rutherford

Motivating example: Shuttle data

```
> data("shuttle", package = "SMPracticals")
> rownames(shuttle) = as.character(rownames(shuttle))
> shuttle[1:4, ]
```

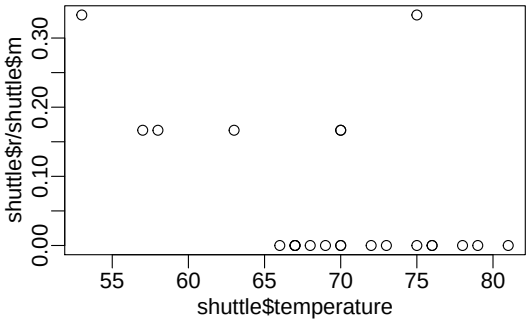
```
  m r temperature pressure
1 6 0          66       50
2 6 1          70       50
3 6 0          69       50
4 6 0          68       50
```

- m: number of rings
- r: number of damaged rings

Questions and models

- Are shuttle rings more likely to get damaged in cold weather?
- Pressure is a confounder
- The data aren't Gaussian
- Binomial distribution r failures from m trials
- Model failure probability as a function of temperature, pressure

```
> plot(shuttle$temperature,
+      shuttle$r/shuttle$m)
```



Motivating example: Fiji birth data

- Fiji Fertility Survey, 1974 opr.princeton.edu/archive/wfs/FJ.aspx
- pbrown.ca/teaching/astwo/data/fiji.RData created from pbrown.ca/teaching/astwo/data/fiji.R

```
> head(fiji)
```

	age	ageMarried	monthsSinceM	failedPreg	pregnancies	children	sons
1	25	18to20	72	0	0	0	0
2	31	15to18	184	0	6	6	2
3	40	15to18	269	0	2	2	1
4	46	15to18	206	0	9	9	6
5	25	18to20	82	0	4	3	1
6	27	20to22	70	0	3	2	1

	firstBirthInterval	residence	literacy	ethnicity
1	60-Inf	rural	yes	fijian
2	12-23	rural	yes	fijian
3	0-7	rural	yes	fijian

Questions and models

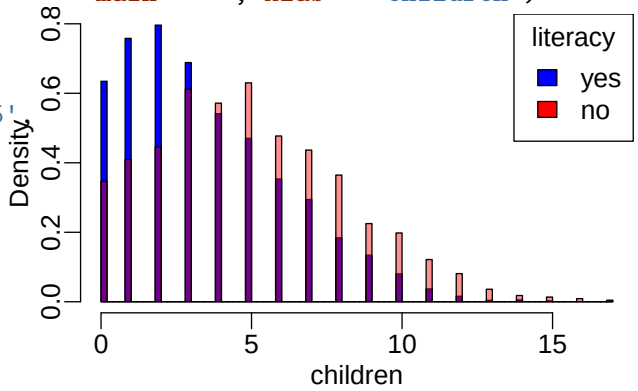
- Do literate women tend to have smaller families?
- Is this only because illiterate women marry earlier?

```
> table(fiji$ageMarried=='0to15',
+   fiji$literacy)
```

	yes	no
FALSE	3483	778
TRUE	323	333

- children $\sim$ Poisson

```
> literate = fiji$literacy == "yes"
> hist(fiji$children[literate], breaks = 100,
+     main = "", xlab = "children")
```



Motivating example: Smoking

- 2014 American National Youth Tobacco Survey

```
> smoke[1:4, c("Age", "Sex", "Race", "RuralUrban", "state",
+   "ever_cigarettes")]
```

```
  Age Sex     Race RuralUrban state ever_cigarettes
1  13   M hispanic      Urban    AZ           FALSE
2  12   F hispanic      Urban    AZ           FALSE
3  14   M   native      Urban    AZ           FALSE
4  13   M hispanic      Urban    AZ           FALSE
```

```
> table(smoke$ever_cigarettes, smoke$Race)
```

```
       white black hispanic asian native pacific
FALSE   7576  2626     4543   844    250      56
TRUE    2243   731     1397   118     83      26
```

America is unusual, both for its obsession with race and for its superb statistics.

from ‘The great melting’, economist.com 9 Jan 2016

```
> table(smoke$Ever_chewing_tobacco_snuf, smoke$Race)

       white black hispanic asian native pacific
  FALSE  8646  3232     5618   947    294      66
  TRUE   1133    98      316    17     36      16
> table(smoke$ever_tobacco_hookah, smoke$Race)

       white black hispanic asian native pacific
  FALSE  8352  2977     4716   858    291      56
  TRUE   1328   299     1054    93     35      23
```

Where we are

Done

- course overview
- stats and data science
- motivating examples

Next

- Scientific writing

References

- Becker, R., J. Chambers, and A. Wilks (1988). *The new S language: a programming environment for data analysis and graphics*. Wadsworth & Brooks/Cole computer science series. Wadsworth & Brooks/Cole Advanced Books & Software. URL: https://search.library.utoronto.ca/details?2959713.
- Breiman, L. (Aug. 2001). "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)". In: *Statist. Sci.* 16.3, pp. 199–231. DOI: 10.1214/ss/1009213726.
- Diggle, P. J. (2015). "Statistics: a data science for the 21st century". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178.4, pp. 793–813. DOI: 10.1111/rssa.12132.
- Donoho, D. (2017). "50 Years of Data Science".
- Fournier, D. A., H. J. Skaug, J. Ancheta, J. Ianelli, A. Magnusson, M. N. Maunder, A. Nielsen, and J. Sibert (2012). "AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models". In: *Optimization Methods and Software* 27.2, pp. 233–249. DOI: 10.1080/10556788.2011.597854.

References I

- Becker, R., J. Chambers, and A. Wilks (1988). *The new S language: a programming environment for data analysis and graphics*. Wadsworth & Brooks/Cole computer science series. Wadsworth & Brooks/Cole Advanced Books & Software. URL: https://search.library.utoronto.ca/details?2959713.
- Breiman, L. (Aug. 2001). "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)". In: *Statist. Sci.* 16.3, pp. 199–231. DOI: 10.1214/ss/1009213726.
- Diggle, P. J. (2015). "Statistics: a data science for the 21st century". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178.4, pp. 793–813. DOI: 10.1111/rssa.12132.
- Donoho, D. (2017). "50 Years of Data Science".

References II

- Fournier, D. A., H. J. Skaug, J. Ancheta, J. Ianelli, A. Magnusson, M. N. Maunder, A. Nielsen, and J. Sibert (2012). "AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models". In: *Optimization Methods and Software* 27.2, pp. 233–249. DOI: 10.1080/10556788.2011.597854.
- Kristensen, K., A. Nielsen, C. Berg, H. Skaug, and B. Bell (2016). "TMB: Automatic Differentiation and Laplace Approximation". In: *Journal of Statistical Software, Articles* 70.5, pp. 1–21. DOI: 10.18637/jss.v070.i05. URL: https://www.jstatsoft.org/v070/i05.
- Rue, H., S. Martino, and N. Chopin (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations". In: *Journal of the Royal Statistical Society B* 71.2, pp. 319–392. DOI: 10.1111/j.1467-9868.2008.00700.x.