Distributions for Parameters

Nancy Reid

April 21, 2017

Morris H. DeGroot Memorial Lecture

CMU STATISTICS
1966–2016
Years

Department of Statistics

Carnegie Mellon University
Dietrich College of Humanities & Social Sciences

Classical Approaches

What are we looking for?

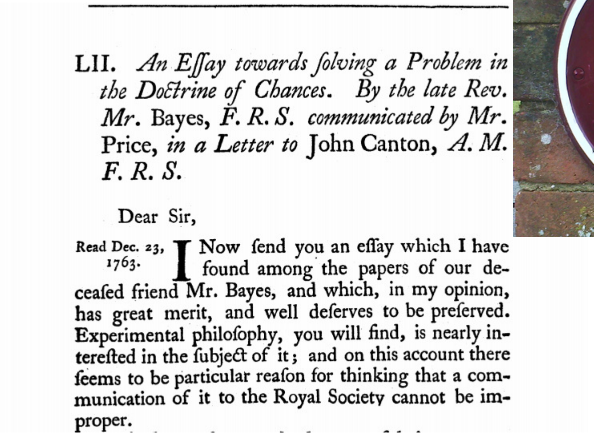
Nature of Probability

Modern Approaches

What's the end goal?

Posterior Distribution

Bayes 1763

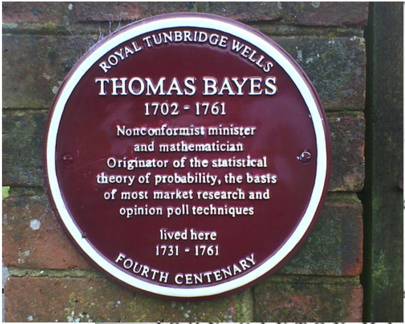


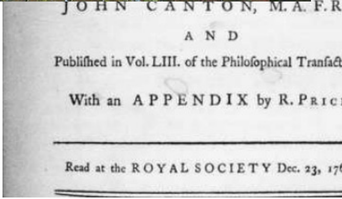
LII. *An Eſſay towards ſolving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

Dear Sir,

Read Dec. 23, 1763.

I Now ſend you an eſſay which I have found among the papers of our deceaſed friend Mr. Bayes, and which, in my opinion, has great merit, and well deſerves to be preſerved. Experimental philoſophy, you will find, is nearly intereſted in the ſubject of it; and on this account there ſeems to be particular reaſon for thinking that a communication of it to the Royal Society cannot be improper.





JOHN CANTON, M.A. F.R

AND

Publiſhed in Vol. LIII. of the Philoſophical Tranſact

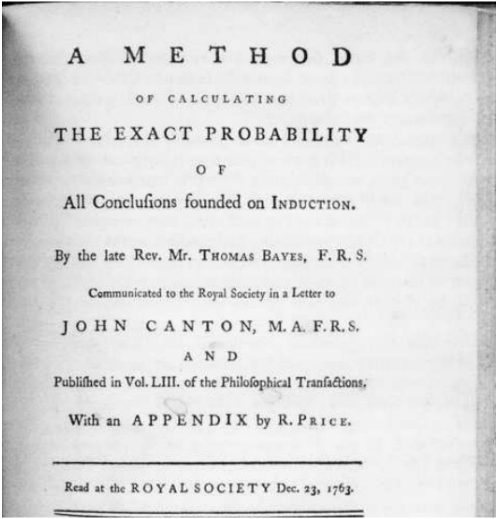
With an APPENDIX by R. PRIC

Read at the ROYAL SOCIETY Dec. 23, 17

Stigler 2013

Posterior Distribution

Bayes 1763



A METHOD

OF CALCULATING

THE EXACT PROBABILITY

OF

All Concluſions founded on INDUCTION.

By the late Rev. Mr. THOMAS BAYES, F. R. S.

Communicated to the Royal Society in a Letter to

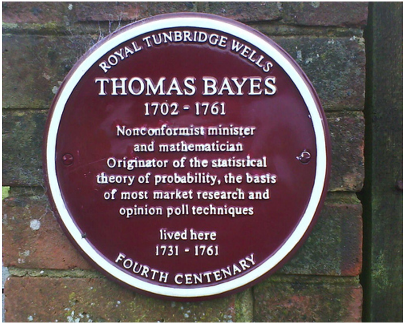
JOHN CANTON, M. A. F. R. S.

AND

Publiſhed in Vol. LIII. of the Philoſophical Tranſactions,

With an APPENDIX by R. PRICE.

Read at the ROYAL SOCIETY Dec. 23, 1763.



$$\pi(\theta \mid y^0) = f(y^0;\theta)\pi(\theta)/m(y^0)$$

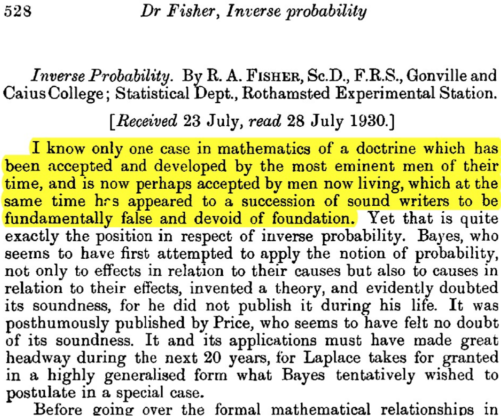
probability distribution for θ

y^0 is fixed

probability comes from $\pi(\theta)$

Fiducial Probability

Fisher 1930



528 *Dr Fisher, Inverse probability*

Inverse Probability. By R. A. Fisher, Sc.D., F.R.S., Gonville and Caius College; Statistical Dept., Rothamsted Experimental Station.

[*Received* 23 July, *read* 28 July 1930.]

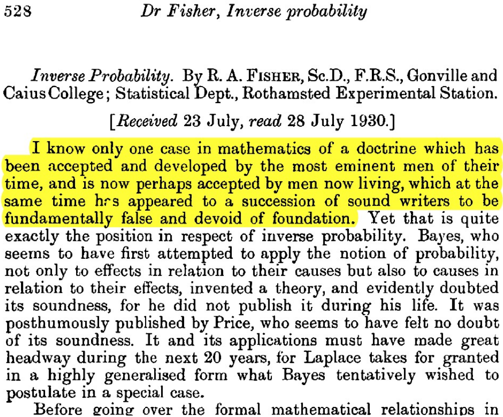
I know only one case in mathematics of a doctrine which has been accepted and developed by the most eminent men of their time, and is now perhaps accepted by men now living, which at the same time has appeared to a succession of sound writers to be fundamentally false and devoid of foundation. Yet that is quite exactly the position in respect of inverse probability. Bayes, who seems to have first attempted to apply the notion of probability, not only to effects in relation to their causes but also to causes in relation to their effects, invented a theory, and evidently doubted its soundness, for he did not publish it during his life. It was posthumously published by Price, who seems to have felt no doubt of its soundness. It and its applications must have made great headway during the next 20 years, for Laplace takes for granted in a highly generalised form what Bayes tentatively wished to postulate in a special case.

Before going over the formal mathematical relationships in

"A small messy man with red hair, a beard and glasses boasting near inch-thick lenses... ... Fisher is remembered as the most significant British statistician of the 20th century" Hampstead Highgate Express, 2013

Fiducial Probability

Fisher 1930



528 *Dr Fisher, Inverse probability*

Inverse Probability. By R. A. Fisher, Sc.D., F.R.S., Gonville and Caius College; Statistical Dept., Rothamsted Experimental Station.

[*Received* 23 July, *read* 28 July 1930.]

I know only one case in mathematics of a doctrine which has been accepted and developed by the most eminent men of their time, and is now perhaps accepted by men now living, which at the same time has appeared to a succession of sound writers to be fundamentally false and devoid of foundation. Yet that is quite exactly the position in respect of inverse probability. Bayes, who seems to have first attempted to apply the notion of probability, not only to effects in relation to their causes but also to causes in relation to their effects, invented a theory, and evidently doubted its soundness, for he did not publish it during his life. It was posthumously published by Price, who seems to have felt no doubt of its soundness. It and its applications must have made great headway during the next 20 years, for Laplace takes for granted in a highly generalised form what Bayes tentatively wished to postulate in a special case.

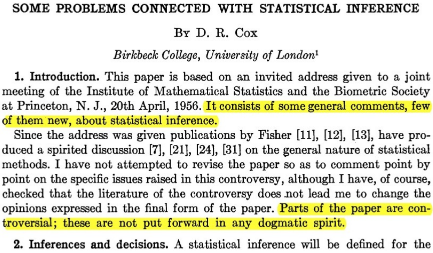
Before going over the formal mathematical relationships in

$$\mathrm{d}f = -\frac{\partial}{\partial\theta} F(T, \theta)\mathrm{d}\theta$$

fiducial probability density for θ, given statistic T
probability comes from (dist'n of) T

Confidence Distribution

Cox 1958; Efron 1993



SOME PROBLEMS CONNECTED WITH STATISTICAL INFERENCE

By D. R. Cox

Birkbeck College, University of London[1]

1. Introduction. This paper is based on an invited address given to a joint meeting of the Institute of Mathematical Statistics and the Biometric Society at Princeton, N. J., 20th April, 1956. It consists of some general comments, few of them new, about statistical inference.

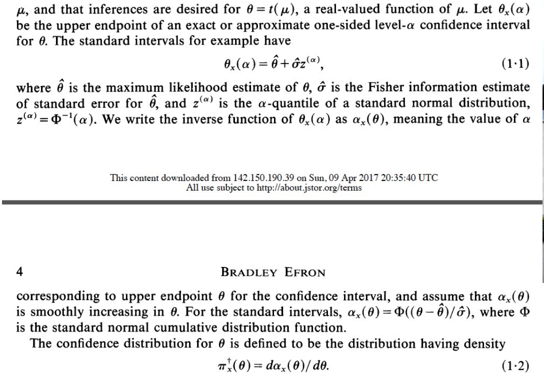
Since the address was given publications by Fisher [11], [12], [13], have produced a spirited discussion [7], [21], [24], [31] on the general nature of statistical methods. I have not attempted to revise the paper so as to comment point by point on the specific issues raised in this controversy, although I have, of course, checked that the literature of the controversy does not lead me to change the opinions expressed in the final form of the paper. Parts of the paper are controversial; these are not put forward in any dogmatic spirit.

2. Inferences and decisions. A statistical inference will be defined for the

- ”Much controversy has centred on the distinction between fiducial and confidence estimation”
- “ ... The fiducial approach leads to a distribution for the unknown parameter”
- “... the method of confidence intervals, as usually formulated, gives only one interval at some preselected level of probability”
- “... in ... simple cases ... there seems no reason why we should not work with confidence distributions for the unknown parameter
- “These can either be defined directly, or ... introduced in terms of the set of all confidence intervals”

Confidence Distribution

Cox 1958; Efron 1993



μ, and that inferences are desired for $\theta = t(\mu)$, a real-valued function of μ. Let $\theta_x(\alpha)$ be the upper endpoint of an exact or approximate one-sided level-α confidence interval for θ. The standard intervals for example have

$$\theta_x(\alpha) = \hat{\theta} + \hat{\sigma} z^{(\alpha)}, \qquad (1{\cdot}1)$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ, $\hat{\sigma}$ is the Fisher information estimate of standard error for $\hat{\theta}$, and $z^{(\alpha)}$ is the α-quantile of a standard normal distribution, $z^{(\alpha)} = \Phi^{-1}(\alpha)$. We write the inverse function of $\theta_x(\alpha)$ as $\alpha_x(\theta)$, meaning the value of α

corresponding to upper endpoint θ for the confidence interval, and assume that $\alpha_x(\theta)$ is smoothly increasing in θ. For the standard intervals, $\alpha_x(\theta) = \Phi((\theta - \hat{\theta})/\hat{\sigma})$, where Φ is the standard normal cumulative distribution function.

The confidence distribution for θ is defined to be the distribution having density

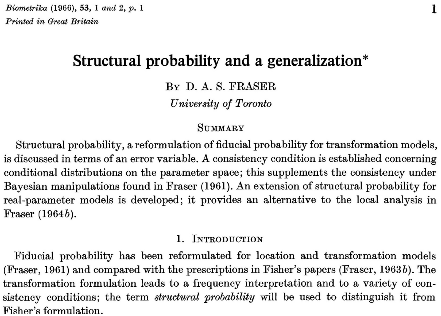
$$\pi_x^{\dagger}(\theta) = d\alpha_x(\theta)/d\theta. \qquad (1{\cdot}2)$$

- “assigns probability 0.05 to θ lying between the upper endpoints of the 0.90 and 0.95 confidence intervals, etc.
- “Of course this is logically incorrect, but it has powerful intuitive appeal”
- “... no nuisance parameters [this] is exactly Fisher’s fiducial distribution”

Seidenfeld 1992; Zabell 1992

Structural Probability

Fraser 1966



Biometrika (1966), **53**, 1 *and* 2, *p.* 1 **1**
Printed in Great Britain

Structural probability and a generalization*

By D. A. S. FRASER
University of Toronto

Summary

Structural probability, a reformulation of fiducial probability for transformation models, is discussed in terms of an error variable. A consistency condition is established concerning conditional distributions on the parameter space; this supplements the consistency under Bayesian manipulations found in Fraser (1961). An extension of structural probability for real-parameter models is developed; it provides an alternative to the local analysis in Fraser (1964*b*).

1. Introduction

Fiducial probability has been reformulated for location and transformation models (Fraser, 1961) and compared with the prescriptions in Fisher's papers (Fraser, 1963*b*). The transformation formulation leads to a frequency interpretation and to a variety of consistency conditions; the term *structural probability* will be used to distinguish it from Fisher's formulation.

- "a re-formulation of fiducial probability for transformation models"
- "This transformation re-formulation leads to a frequency interpretation"
- a change in the parameter value can be offset by a change in the sample

$$y \to y + a; \theta \to \theta - a$$

- a local location version leads to:

$$\mathrm{d}f = -\frac{\partial}{\partial\theta}F(y,\theta)\mathrm{d}\theta = -\frac{\partial}{\partial\theta}F(y,\theta)\frac{f(y^0,\theta)}{f(y^0,\theta)} = \overbrace{f(y^0,\theta)}^{\text{Likelihood}}\left.\frac{dy}{d\theta}\right|_{y^0}$$

Significance Function

Fraser 1991

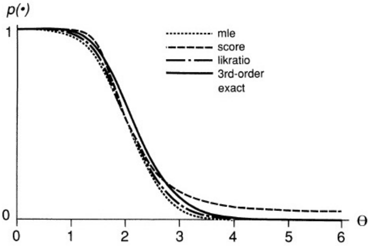
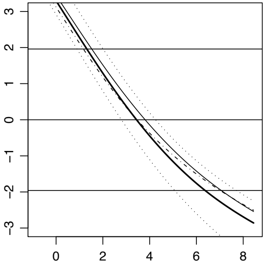


Figure 5. The Standardized Maximum Likelihood Estimate, Standardized Score, and Signed Likelihood Ratio Produce Three Approximations for the Significance Function. Model: location log gamma(3); data: $y^0 = 3.14$.



- “from likelihood to significance”
- “significance records probability left of the observed data point”

  ”likelihood records probability at the observed data point”
- the significance function is a plot of this probability

  as a function of θ
- “the full spectrum of confidence intervals is obtained ... suggesting the alternate name confidence distribution function

Why do we want distributions on parameters?

- inference is intuitive
- combines easily with decision theory
- de-emphasizes point estimation and arbitrary cut-offs

- Example:
  $n = 10, \bar{y} = 1.58, s = 1.23, s/\sqrt{n} = 0.39, t(\mu) = \sqrt{n}(\bar{y} - \mu)/s$
- If μ is the true value, then $\mathrm{pr}\{t_{\alpha/2} \leq t(\mu) \leq t_{1-\alpha/2}\} = 1 - \alpha$
- pivot on t to obtain
  $$(1-\alpha)CI : \{\bar{y} - t_{1-\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{y} + t_{\alpha/2}\frac{s}{\sqrt{n}}\} = (0.70, 2.46)$$

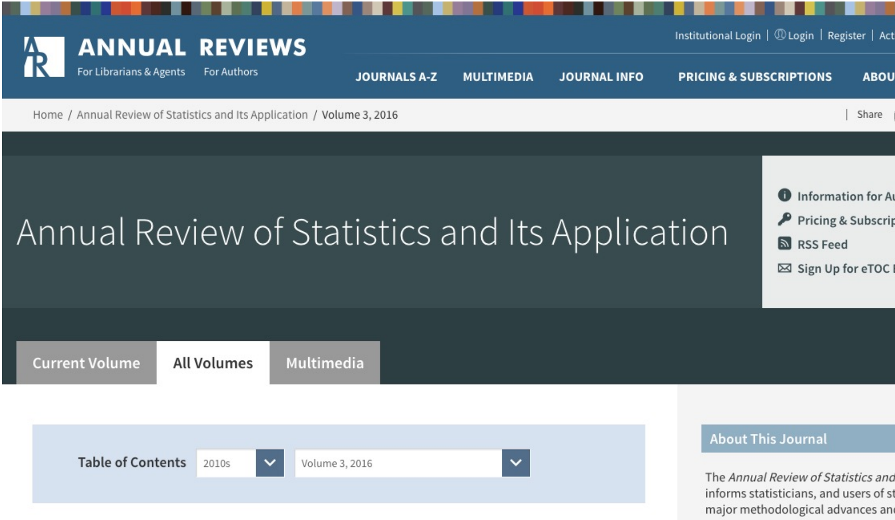
- "it's tempting to conclude that μ is more likely to be near the middle of this interval, and if outside, not very far outside"

Cox 2006

Why not go Bayes?

Example: League Tables for Hospital Comparisons

Normand, Ash, Fienberg, Stukel, Utts, Louis ARSIA V3



Annual Review of Statistics and Its Application

The *Annual Review of Statistics and* informs statisticians, and users of st major methodological advances an

League Tables

Normand et al. 2016

- CMS uses a hierarchical model to model death risks to accommodate patient-level variation in outcome, patient-level risk, and hospital-level variation
- Model: Y_{ij} a binary outcome (death) for patient j, with risk factors x_{ij}, at hospital i, with n_i cases

$$\begin{aligned} Y_{ij} \mid \beta_{0i}, \alpha, x_{ij} &\sim \text{Bern}\{p_{ij}\}, \\ \text{logit}(p_{ij}) &= \beta_{0i} + \alpha^{\mathrm{T}} x_{ij}, \\ \beta_{0i} \mid \mu, \tau^2 &\sim N(\mu, \tau^2) \end{aligned}$$

- $\text{SMR}_i = \dfrac{\sum_{j=1}^{n_i} \text{ED}_i(\beta_{0i}, x_{ij}, \alpha, \mu, \tau^2)}{\sum_{j=1}^{n_i} \text{ED}_i(x_{ij}, \mu, \alpha, \tau^2)}$
- The numerator integrates over the posterior distribution of β_{0i}, and the denominator integrates over the prior distribution of β_{0i}
- Adjusting each hospital's outcomes for its size and case mix

Nature of Probability

Cox 2006; R & Cox 2015; Zabell, 1992

- probability to describe physical haphazard variability
  - probabilities represent features of the “real” world in somewhat idealized form
  - subject to empirical test and improvement
  - conclusions of statistical analysis expressed in terms of interpretable parameters
  - enhanced understanding of the data generating process

- probability to describe the uncertainty of knowledge
  - measures rational, supposedly impersonal, degree of belief, given relevant information Jeffreys, 1939,1961
  - measures a particular person’s degree of belief, subject typically to some constraints of self-consistency

    F.P. Ramsey, 1926; de Finetti, 1937; Savage, 1956

  - often linked with personal decision making necessarily?

... nature of probability

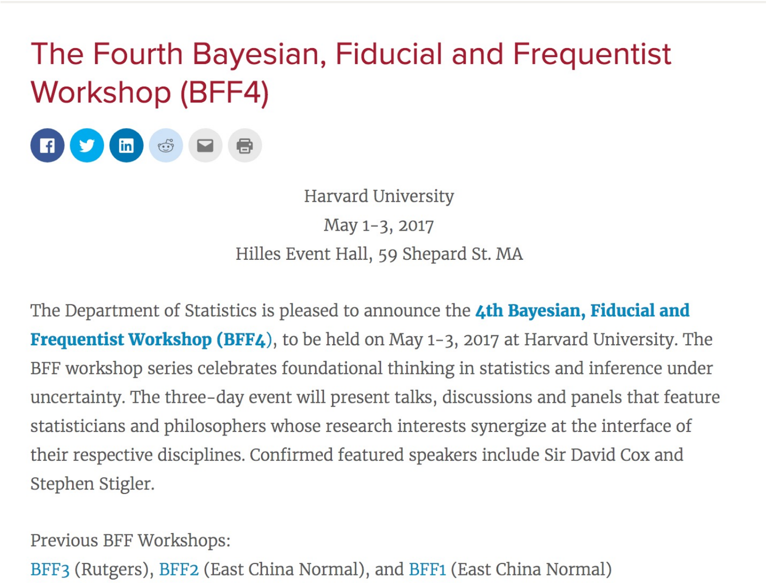
- Bayes posterior describes uncertainty of knowledge
- probability comes from the prior
- or from the model, cf. hospital league tables

- confidence intervals or *p*-values refer to empirical probabilities

- in what sense are confidence distribution functions, significance functions, structural or fiducial probabilities to be interpreted?

- empirically? degree of belief?
- literature is not very clear imho
- we may avoid the need for a different version of probability by appeal to a notion of calibration

What goes around ...



The Fourth Bayesian, Fiducial and Frequentist Workshop (BFF4)

Harvard University
May 1-3, 2017
Hilles Event Hall, 59 Shepard St. MA

The Department of Statistics is pleased to announce the **4th Bayesian, Fiducial and Frequentist Workshop (BFF4)**, to be held on May 1-3, 2017 at Harvard University. The BFF workshop series celebrates foundational thinking in statistics and inference under uncertainty. The three-day event will present talks, discussions and panels that feature statisticians and philosophers whose research interests synergize at the interface of their respective disciplines. Confirmed featured speakers include Sir David Cox and Stephen Stigler.

Previous BFF Workshops:
BFF3 (Rutgers), BFF2 (East China Normal), and BFF1 (East China Normal)

What goes around ...

BFF1,2: “facilitate the exchange of recent research developments in Bayesian, fiducial and frequentist methodology, concerning statistical foundations”

BFF3: “re-examine the foundations of statistical inferences; develop links to bridge gaps among different statistical paradigms”

BFF4: “celebrates foundational thinking in statistics and inference under uncertainty”

What's old is new

- posterior distribution
- fiducial probability
- confidence distribution
- structural probability

- objective Bayes
- generalized fiducial inference
- confidence distributions and confidence curves
- approximate significance functions

Objective Bayes

Berger, BFF4, e.g.

- noninformative, default, matching, reference, ... priors
- we may avoid the need for a different version of probability by appeal to a notion of calibration

Cox 2006, R & Cox 2015

- as with other measuring devices
  within this scheme of repetition, probability is defined as a hypothetical frequency
- it is unacceptable if a procedure yielding high-probability regions in some non-frequency sense are poorly calibrated
- such procedures, used repeatedly, give misleading conclusions

Bayesian Analysis, V1(3) 2006

... objective Bayes

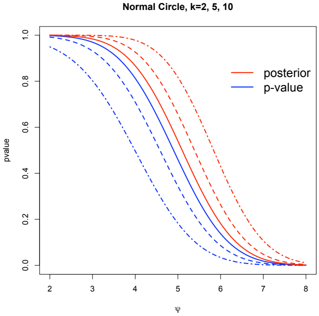
- pragmatic solution as a starting point
- some versions may not be correctly calibrated
- requires checking in each example
- calibrated versions must be targetted on the parameter of interest
- only in very special cases can calibration be achieved for more than one parameter in the model, from the same prior
- the simplicity of a fully Bayesian approach to inference is lost

Gelman 2008; PPM LW

Example

Stein, 1959

- $y_i \sim N(\mu_i, 1/n), \quad i = 1, \ldots, k; \quad \pi(\mu_i) \propto 1$
- posterior distribution of $a^{\mathrm{T}}\mu$ is well-calibrated
- marginal posterior distribution of $||\mu||$ is not
- discrepancy is a function of $\dfrac{k-1}{||\mu||\sqrt{n}}$



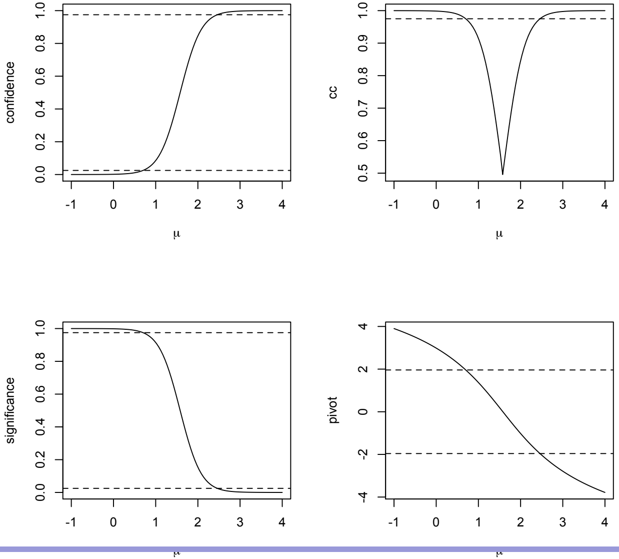
Confidence Distribution

Xie & Singh; Hjort & Schweder

- any function $H : \mathcal{Y} \times \Theta \to (0,1)$ which is
- a cumulative distribution function of θ for any $y \in \mathcal{Y}$
- has correct coverage: $H(Y,\theta) \sim U(0,1)$ $Y \sim f(\cdot;\theta)$
- CDs, or approximate CDs, are readily obtained from pivotal quantitites
- pivotal quantity: $g(y,\theta)$ with sampling distribution known $\sqrt{n}(\bar{y}-\mu)/s$
- sufficiently general to encompass bootstrap distribution and many standard likelihood quantities
- recent examples include robust meta-analysis and identification of change-points

Xie et al. 2011; Cunen et al. 2017; Hannig & Xie 2012

... confidence distribution



Generalized Fiducial

Hannig et al 2016

- Fisher: $g(Y, \theta)$ has a known distribution; invert this to create distribution for θ when y^0 is obtained
- Fraser: use data-generating equation to make the inversion more direct, e.g. $Y_i = \mu + \sigma e_i, \quad e_i = (y_i^0 - \mu)/\sigma$

- Hannig et al. $Y = G(U, \theta), \quad U$ has known distribution
- suppose we can invert this for any y^0: $\theta = Q_{y^0}(U)$
- fiducial distribution of θ is $Q_{y^0}(U^*)$ $\quad$ U^* independent copy of U

- inverse only exists if θ and Y have same dimension
- might get this by reduction of a sample to sufficient statistics
- more generally, some conditional argument seems to be required

Significance Function

Fraser & R, ...

- current solutions based on asymptotic arguments
- that relies on a location model approximation
- which gives appropriate conditioning
- and an exponential model approximation
- which can be computed accurately using saddlepoint approximation
- combination of structural model and likelihood asymptotics
- leads for example to construction of default priors

Fraser et al 2010

What’s the end goal?

- Applications – something that works
  - gives ‘sensible’ answers
  - not too sensitive to model assumptions
  - computable in reasonable time
  - provides interpretable parameters

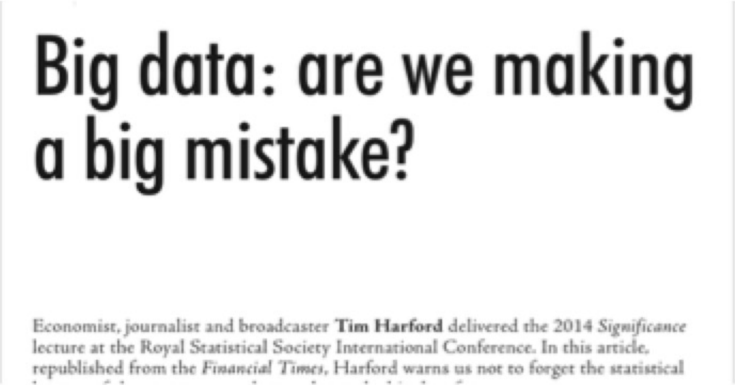
- Foundations – peeling back the layers
  - what does ’works’ mean?
  - what probability do we mean
  - ‘Goldilocks’ conditioning
  - how does this impact applied work?

Meng & Liu, 2016

Role of Foundations

- avoid apparent discoveries based on spurious patterns
- to shed light on the structure of the problem
- calibrated inferences about interpretable parameters
- realistic assessment of precision
- understanding when/why methods work/fail

Some warning signs

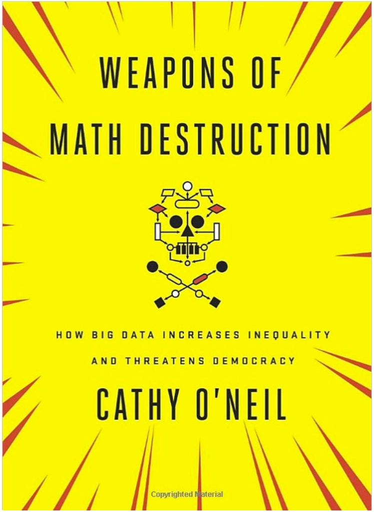


Big data: are we making a big mistake?

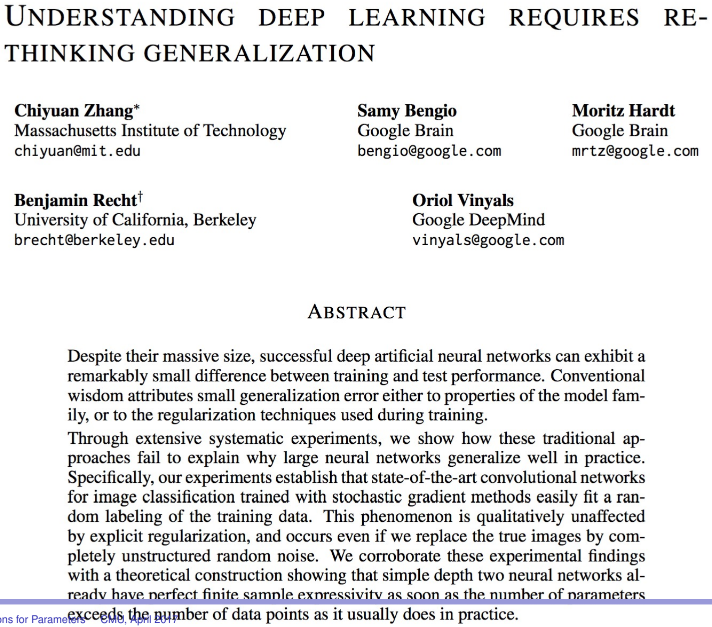
Economist, journalist and broadcaster **Tim Harford** delivered the 2014 *Significance* lecture at the Royal Statistical Society International Conference. In this article, republished from the *Financial Times*, Harford warns us not to forget the statistical

“Big data” has arrived, but big insights have not

Some warning signs



Some warning signs



UNDERSTANDING DEEP LEARNING REQUIRES RETHINKING GENERALIZATION

Chiyuan Zhang[*]
Massachusetts Institute of Technology
chiyuan@mit.edu

Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

Benjamin Recht[†]
University of California, Berkeley
brecht@berkeley.edu

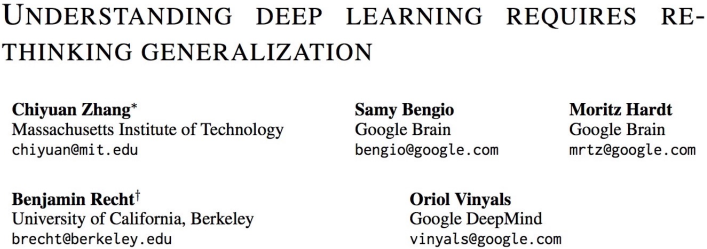
Oriol Vinyals
Google DeepMind
vinyals@google.com

ABSTRACT

Despite their massive size, successful deep artificial neural networks can exhibit a remarkably small difference between training and test performance. Conventional wisdom attributes small generalization error either to properties of the model family, or to the regularization techniques used during training.

Through extensive systematic experiments, we show how these traditional approaches fail to explain why large neural networks generalize well in practice. Specifically, our experiments establish that state-of-the-art convolutional networks for image classification trained with stochastic gradient methods easily fit a random labeling of the training data. This phenomenon is qualitatively unaffected by explicit regularization, and occurs even if we replace the true images by completely unstructured random noise. We corroborate these experimental findings with a theoretical construction showing that simple depth two neural networks already have perfect finite sample expressivity as soon as the number of parameters exceeds the number of data points as it usually does in practice.

Some warning signs



UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

Chiyuan Zhang*
Massachusetts Institute of Technology
chiyuan@mit.edu

Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

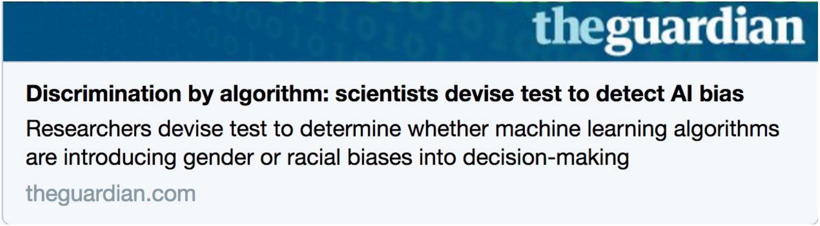
Benjamin Recht[†]
University of California, Berkeley
brecht@berkeley.edu

Oriol Vinyals
Google DeepMind
vinyals@google.com

neural networks easily fit random labels”

“these observations rule out all of VC-dimension, Rademacher complexity, and uniform stability as possible explanations for the generalization performance of state-of-the-art neural networks.”

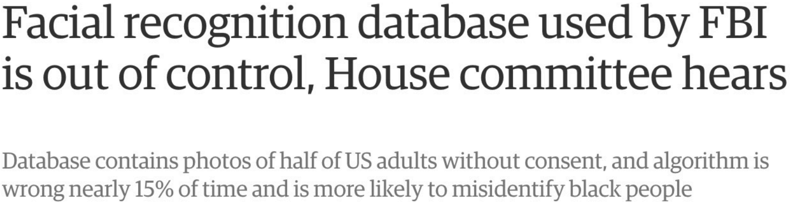
Some warning signs



Discrimination by algorithm: scientists devise test to detect AI bias

Researchers devise test to determine whether machine learning algorithms are introducing gender or racial biases into decision-making

theguardian.com



Facial recognition database used by FBI is out of control, House committee hears

Database contains photos of half of US adults without consent, and algorithm is wrong nearly 15% of time and is more likely to misidentify black people

Summary

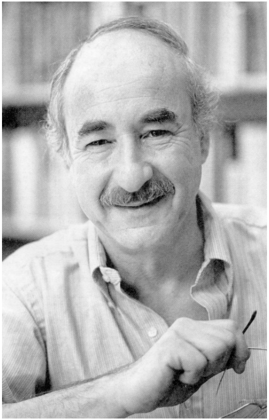
- Bayes, fiducial, structural, confidence

- BFF 1 - 4: Develop links to bridge gaps among different statistical paradigms

- targetting parameters
- limit distributions
- calibration in repeated sampling
- relevant repetitions for the data at hand

NR: Why is conditional inference so hard?
DRC: I expect we're all missing something, but I don't know what it is

StatSci Interview 1996



THANK YOU!

References

Cox, D.R. (1958). *Ann. Math. Statist.*

Cox, D.R. (2006). *Principles of Statistical Inference.*

Cunen et al. (2017). *J. Statist. Plann. Infer.*

Efron, B. (1993). *Biometrika*

Fisher, R.A. (1930). *Proc. Cam. Phil. Soc.*

Fraser, D.A.S. (1966). *Biometrika*

Fraser, D.A.S. (1991). *J. Amer. Statist. Assoc.*

Fraser, D.A.S. et al. (2010). *J. R. Statist. Soc. B*

Fraser, D.A.S. and Reid, N. (1993). *Statist. Sinica*

Gelman, A. (2008). *Ann. Appl. Statist.*

Hannig, J. and Xie, M. (2012). *Elect. J. Statist.*

Hannig, J. et al. (2016). *J. Amer. Statist. Assoc.*

Hjort, N. and Schweder, T. *Confidence, Likelihood, Probability: Inference with Confidence Distributions*

Meng, X.-L. and Liu, K. (2016). *Ann. Rev. Stat. and its Applic.*

Norman, S. et al. (2016). *Ann. Rev. Stat. and its Applic.*

Reid, N. and Cox, D.R. (2015). *Intern. Statist. Rev.*

Stein, C. (1959). *Ann. Math. Statist.*

Stigler, S. (2013). *Statistical Science*

Xie, M. and Singh, K. (2013) *Internat. Statist. Rev.*

Xie, M. et al. (2011). *J. Amer. Statist. Assoc.*

Zabell, X. (1992). *Statistical Science*