Distributions for Parameters

Nancy Reid
University of Toronto

O'Bayes 2019, University of Warwick

July 1 2019

Classical Approaches: A Look Way Back

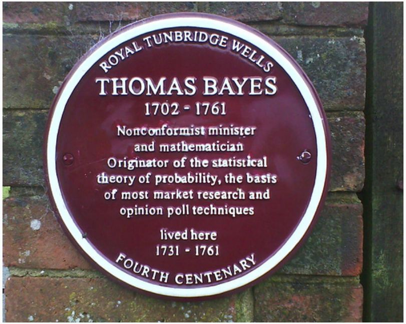
Posterior Distribution

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.

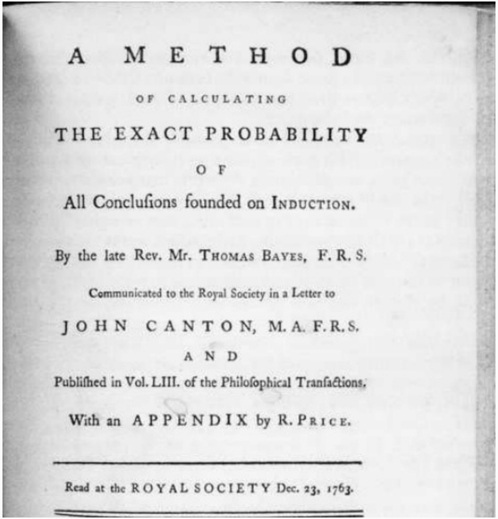


A METHOD

OF CALCULATING

THE EXACT PROBABILITY

Posterior Distribution



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

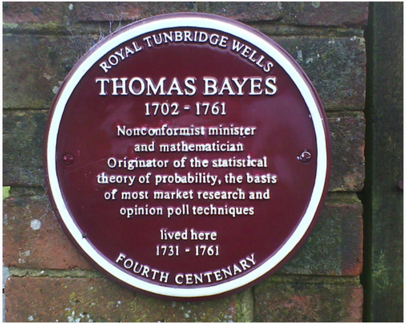
JOHN CANTON, M. A. F. R. S.

AND

Publiſhed in Vol. LIII. of the Philoſophical Tranſactions.

With an APPENDIX by R. PRICE.

Read at the ROYAL SOCIETY Dec. 23, 1763.

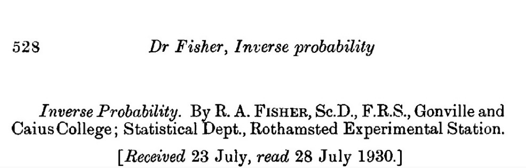


$$\pi(\theta \mid y^o) = f(y^o; \theta)\pi(\theta)/m(y^o)$$

probability distribution for θ

y^o is fixed

probability comes from $\pi(\theta)$



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.]

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

fiducial probability density for θ, given statistic T

probability comes from (dist'n of) T

"It is not to be lightly supposed that men of the mental calibre of Laplace and Gauss ... could fall into error on a question of prime theoretical importance, without an uncommonly good reason"

Confidence Distribution

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

The idea of obtaining Bayesian results from confidence intervals goes back at least to Fisher's work on fiducial inference in the 1930's. Suppose that a data set x is observed from a parametric family of densities $g_\mu(x)$, depending on an unknown parameter vector 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

$$\theta_x(\alpha) = \hat{\theta} + \hat{\sigma} z^{(\alpha)}, \qquad (1\cdot1)$$

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

$$\pi_x^\dagger(\theta) = d\alpha_x(\theta)/d\theta. \qquad (1\cdot2)$$

We shall call (1·2) the confidence density. This distribution assigns probability 0·05 to θ lying between the upper endpoints of the 0·90 and 0·95 confidence intervals, etc. Of

- "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

Biometrika (1966), **53**, 1 *and* 2, *p.* 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 $\quad y \to y + a; \theta \to \theta - a$
- a local location version leads to:

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

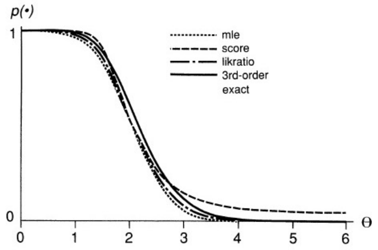
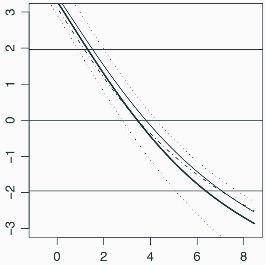


Figure 5. The Standardized Maximum Likelihood Estimate, Standardized Score, and Signed Likelihood Ratio Produce Three Approximations for the Significance Function. Model: location log gamma(3); data: $y^0 = 3.14$.



- “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 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

- "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

- "assigns probability 0.05 to θ lying between the upper endpoints of the 0.90 and 0.95 confidence intervals, etc."

Efron 1993

- all inference statements become probability statements about unknowns hmm...

That was then

- posterior distribution $\pi(\theta \mid y)$
- fiducial distribution $\mathrm{d}f = -F_\theta(y, \theta)\mathrm{d}\theta$
- confidence distribution $\pi_y(\theta) = \mathrm{d}\alpha_y(\theta)/\mathrm{d}\theta$
- significance function $p(\theta) = \Pr(Y \leq y^o; \theta)$
- belief functions

Dempster 1966; Shafer 1976

BFF one to six

BFF 1 – 6

BFF1,2: “facilitate the exchange of recent research developments in Bayesian, fiducial and frequentist methodology, concerning statistical foundations” 2014,5

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”

BFF5: “… and foundations of data science”

BFF6: “… and model uncertainty”

BFF7: “… methodological, computational, and ethical principles of data science”


BFF
BEERFOODFOREVER

… BFF 1 – 6

- posterior distribution
- fiducial probability
- confidence distribution
- structural probability
- significance function
- belief functions
- model misspecification

- objective Bayes
- generalized fiducial inference
  Hannig; Taraldsen
- confidence distributions and confidence curves
  Hjort, Schweder, Xie
- approximate significance functions
  Brazzale et al; Fraser & R
- inferential models
  Martin & Liu
- generalized Bayes
  pseudo-posterior, safe Bayes
  tempered likelihood

What has changed?

computation

model complexity

model dimension

data

science

From geeky to cool: Statistics is Berkeley’s fastest-growing major

… BFF 1 – 6

- posterior distribution
- fiducial probability
- confidence distribution
- structural probability
- significance function
- belief functions
- model misspecification

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

pseudo-posterior, safe Bayes

tempered likelihood

Objective Bayes

Objective Bayes

- 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

... objective Bayes

- the simplicity of a fully Bayesian approach to inference is lost — Gelman 2008
- meaning?
- $$\pi(\psi \mid y) = \int_{\psi(\theta)=\psi} \pi(\theta \mid y) d\theta, \quad \text{for any } \psi : \Theta \searrow \Psi$$

  lower dimension
- this is the step where the prior can have unexpected influence
  — flat priors can be a disaster — objective Bayes fails
- as does fiducial — Dawid, BFF 6
- theory of confidence distributions gets around this by constructing a CD separately for each parameter of interest — Hjort and Schweder , 2016, focus parameter
- theory of significance functions targets parameter of interest using a higher order pivotal — Fraser 2017 ARSIA

Example

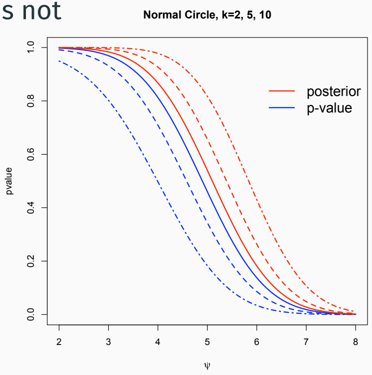
- $y_i \sim N(\mu_i, 1/n), \quad i = 1, \ldots, k; \quad \pi(\mu_i) \propto 1$
- posterior distribution of $a^{\mathrm{T}}\mu$ is well-calibrated
- marginal posterior distribution of $\psi = ||\mu||$ is not
- discrepancy is a function of

$$\frac{k-1}{\psi\sqrt{n}}$$

- global-local shrinkage priors (horseshoe) shrink the posterior in the right direction

Bhadra et al., Bka, 2016

- reference and targetted priors do the same



... objective Bayes

- calibrated versions must be targetted on the parameter of interest

- penalized complexity priors: separate the parameters 'as much as possible', put independent priors adapted to the model — Rue, OBayes 2019

- new class of objective priors — Walker, OBayes 2019
  - doesn't depend on the model
  - doesn't reduce to Jeffreys' prior in $\mathbb{R}^1$
  - can't be calibrated
  - can't be marginalized (?)

- objective Bayes is not Bayes — Lewis, OBayes 2019
  - problems with updating sequentially

Generalized Bayes

Sparse normal means

- also called sequence model — Castillo et al. 2015
- $y_i \sim N(\theta_i, 1), i = 1, \dots, n; \quad \theta$ 'sparse' — more generally $y_i \sim N(x_i^{\mathrm{T}}\beta, \sigma^2)$; β 'sparse'
    - informative prior on $|S|$
    - uniform prior on S given $|S|$
    - normal prior on θ_S, given S : $\quad N_{|S|}(Y_S, \gamma^{-1} I_{|S|})$ — $Y_S = \{y_i, i \in S\}$
- doesn't use Normal likelihood
- uses instead

$$\pi(\cdot \mid y) \propto \{N_{|S|}(\theta_S, \gamma^{-1} I_{|S|})\}^{\alpha} \pi(\theta, |S|, S)$$

$\alpha < 1$ to be determined

$$\pi(\theta \mid y) \propto \{L_n(\theta; y)\}^{\alpha} \pi(\theta)$$

generalized Bayes; Gibbs posterior, corrected likelihood

- another approach has prior on β as product of Laplace densities; no tempering needed for convergence of $\pi_m(\theta \mid y)$ — Castillo et al 2015

Aside: composite likelihood

- multivariate response $y \sim f(\cdot;\theta)$
- likelihood function $L(\theta;y) \propto \prod_{i=1}^{n} f(y_i;\theta)$
- composite likelihood function $cL(\theta;y) \propto \prod_{i=1}^{n}\prod_{r<s} f_2(y_{ir},y_{is};\theta)$
- not a real likelihood — probability of an observable random vector
- doesn't give a real posterior — misspecified model
- doesn't give usual asymptotics, e.g. $2\log\{cL(\hat{\theta}_{CL};y)/cL(\theta);y\} \xrightarrow{d} \Sigma_j \lambda_j \chi^2_{1j}$
- proposal: compute 'posterior' as

$$\pi(\theta \mid y) \propto \{cL(\theta;y)\}^{1/\tilde{\lambda}_n}\,\pi(\theta)$$

- $\tilde{\lambda}$ related to information matrices for $\log cL(\theta;y)$ — Pauli, Racugno, Ventura 2011
- patches up model misspecification caused by using $cL(\theta;y)$ instead of $L(\theta;y)$

- Dunson and Miller propose a c-posterior

$$\pi^c(\theta \mid y) \propto \{L(\theta; y)\}^{a/(a+n)}$$

a a hyper-parameter

- based on modelling misspecification via a distance between observations $y_1, \ldots, y_n$ and 'ideal' observations $Y_1, \ldots, Y_n$
- see also 'coverage inducing priors' Müller & Norets 2016
- Grünewald and van Ommen refer to 'generalized posterior'

$$\pi(\theta \mid y) \propto \{L(\theta; y)\}^{\eta} \pi(\theta), \eta < 1$$

- argues that this can be interpreted as 'ordinary Bayes' with a different model

Walker and Hjort 2002; data-dependent prior

Comparisons

Comparisons: conditioning

• objective Bayes	• yes
• generalized fiducial inference	• yes and no JASA ’16
• confidence distributions and confidence curves	• needs to be built in ahead of time
• approximate significance functions	• yes; via approximate location model
• inferential models	• needs to be built in ahead of time
• generalized Bayes	• yes, as with objective Bayes data is fixed

Comparisons: Eliminating Nuisance Parameters

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

- marginalization

  rarely works ...
- depends on the problem

  ?
- use profile log-likelihood, or similar

  focus parameter
- marginalization

  via Laplace approximation
- marginalization

  invoked ahead of time
- ?? probably not easy

Comparisons: Calibration

• objective Bayes	• often
• generalized fiducial inference	• yes
• confidence distributions and confidence curves	• typically approximate
• approximate significance functions	• typically approximate
• inferential models	• yes
• generalized Bayes	• ?? possibly

Comparisons: Nature of Probability

Bayes / objective Bayes	epistemic / empirical
generalized fiducial inference	empirical
confidence distributions and confidence curves	empirical but not prescriptive
approximate significance functions	empirical
inferential models	epistemic
generalized Bayes	not

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 Meng & Liu, 2016
  - how does this impact applied work?

Role of Foundations

Cox & R, 2015

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

Are we making progress?

objective Bayes

Larry W: “the perpetual motion machine of Bayesian inference”

confidence distributions

Min-ge, Regina: “everything fits”
Nils: “CDs are the ‘gold standard’ ”

generalized fiducial

Jan: “bring it on ... I’ll figure it out”

inferential models

Ryan: “it’s the only solution”
Chuanhai: “ it might take 100 years”

significance functions

Don: “it’s the best solution ...
you can’t solve everything at once”

… are we making progress?

confidence distributions

Min-ge, Regina: “everything fits”
Nils: “CDs are the ‘gold standard’ ”

generalized fiducial

Jan: “bring it on … I’ll figure it out”

inferential models

Ryan: “it’s the only solution”
Chuanhai: “ it might take 100 years”

significance functions

Don: “it’s the best solution …
you can’t solve everything at once”

generalized Bayes

Peter: “ideally develop … a general theory of substitution likelihoods; pseudo(composite) likelihood, rank-based likelihood, would become special cases”

Summary

- Bayes, fiducial, structural, confidence, belief

- BFF 1 - 6: Develop links to bridge gaps among different statistical paradigms

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

- complex models, high-dimensional parameters

Rue, OBayes 2019

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

References

Bayesian Analysis V1 Issue 3.

Castillo et al. (2015). Ann. Statist.

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

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

Dempster, A. (1966). Ann. Math. Statist.

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. (2017). Ann. Rev. Statist. Applic.

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.

Grünewald, P. and van Ommen, T. (2017) Bayesian Anal.

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

Hjort, N. and Schweder, T. (2016). Confidence, Likelihood, Probability: Inference with Confidence Distributions

Martin, R. and Liu, C. (2015). Inferential Models

Martin, R. and Ng, B. (2018). `http://www.researchers.one/article/2018-12-6`

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

Miller, J. and Dunson, D.

Müller, U.K. and Norets, A. (2016). JASA

Pauli, F., Raccugno, W., and Ventura, L. (2011). Statist. Sinica Reid, N. and Cox, D.R. (2015). Intern. Statist. Rev.

Seidenfeld, T. (1992). Statist. Sci.

Shafer, G. (1976). A Mathematical Theory of Evidence.

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

Taraldsen, G. and Lindqvist, B.H. (2018). JSPI

Walker, S. and Hjort, N. (2002) JRSS B

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

Zabell, X. (1992). Statistical Science