

Principles of Statistical Inference

Nancy Reid and David Cox

August 30, 2013

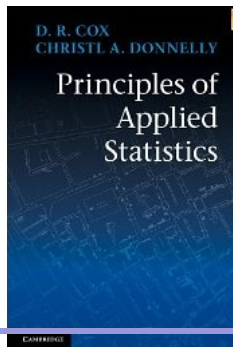
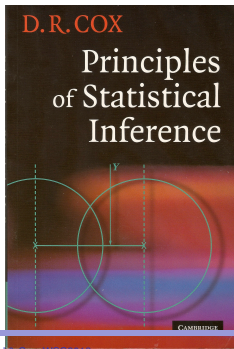


Introduction

- ▶ Statistics **needs** a healthy interplay between theory and applications
 - ▶ theory meaning **Foundations**, rather than theoretical analysis of specific techniques
- ▶ **Foundations?** suggests a solid base, on which rests a large structure
 - ▶ must be continually tested against new applications
- ▶ The notion that this can be captured in a simple framework, much less a set of mathematical axioms, seems dangerously naive Fisher 1956
- ▶ Foundations in Statistics depend on, and must be **tested and revised**, in the light of experience
- ▶ and assessed by **relevance** to the very wide variety of contexts in which statistical considerations arise

... introduction

- ▶ A formal theory of inference is just a small part of the challenges of statistical work
- ▶ Equally important are aspects of
 - ▶ study design
 - ▶ types of measurement
 - ▶ formulation of research questions
 - ▶ connecting these questions to statistical models
 - ▶ ...



Outline

1. the role of probability
2. some classical principles: sufficiency, ancillarity, likelihood
3. some less classical compromises: pivotals, strong matching, asymptotics, bootstrap
4. some thoughts on the future

Role of probability

- ▶ central to most formulations of statistical issues
- ▶ but not all, e.g. algorithmic approaches popular especially in machine learning Breiman 1999
- ▶ theory of probability has been liberated from discussions of its meaning via Kolmogorov's axioms
- ▶ except possibly the modification needed for quantum mechanics, and notions of upper and lower probability
- ▶ statisticians do not have this luxury!
- ▶ we continue to be engaged by the distinction between
 - ▶ probability representing physical haphazard variability Jeffreys – “chances”
 - ▶ probability encapsulating, directly or indirectly, aspects of the uncertainty of knowledge

Probability and empirical variability

four related, but different, approaches

1. the data are regarded as a random sample from a hypothetical infinite population; frequencies within this are probabilities; some aspect of these represent the target of inference
2. the data form part of a long real or somewhat hypothetical process of repetition under constant conditions; limiting frequencies in this repetition are the probabilities of interest; some aspect ...
3. either 1., 2. or both, plus an explicit, idealized, description of the physical, biological, ... processes that generated the data
4. either 1., 2. or both, used only to describe the randomization in experimental design or in sampling an existing population; leading to the so-called design approach to analysis

... empirical variability

- ▶ probabilities represent features of the “real” world, of course in somewhat idealized form, and, given suitable data, are subject to empirical test and improvement
- ▶ conclusions of statistical analysis are to be expressed in terms of interpretable parameters describing such a probabilistic representation of the system under study
- ▶ enhanced understanding of the data generating process as in epidemics, for example

Probability as uncertain knowledge

probability as measuring strength of belief in some uncertain proposition is sharply different; how do we address this

1. consider that probability measures rational, supposedly impersonal, degree of belief, given relevant information

Jeffreys 1939 1961

2. consider that probability 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

seems intimately linked with personal decision making

3. avoid the need for a different version of probability by appeal to a notion of calibration

... uncertain knowledge

- ▶ we may avoid the need for a different version of probability by appeal to a notion of calibration
- ▶ as assessed by behaviour of a procedure under hypothetical repetition
- ▶ as with other measuring devices within this scheme of repetition, probability is defined as a hypothetical frequency
- ▶ the precise specification of the assessment process may requiring some notion of conditioning

Good 1949

- ▶ the formal accept-reject paradigm of Neyman-Pearson theory would be an instance of decision analysis and as such outside the immediate discussion

A brief assessment

1. even if a frequency view of probability is not used directly as a basis for inference it is unacceptable if
 - ▶ a procedure using probability in another sense is poorly calibrated
 - ▶ such a procedure, used repeatedly, gives misleading conclusions

Bayesian Analysis 1(3) 2006; Wasserman *BA* 3(3) 2008

2. standard accounts of probability assume total ordering of probabilities
 - ▶ can we regard a probability, $p = 0.3$, say, found from careful investigation of a real-world effect as equivalent to the same p derived from personal judgment, based on scant or no direct evidence?

... assessment

3. a great attraction of Bayesian arguments is that all calculations are by the rules of probability theory
 - ▶ however, personalistic approaches merge seamlessly what may be highly personal assessments with evidence from data, possibly collected with great care
 - ▶ this is surely unacceptable for the careful discussion of the meaning of data and the presentation of those conclusions in the scientific literature
 - ▶ even if essential for personal decision making
 - ▶ this is in no way to deny the role of personal judgement and experience in interpreting data; it is the merging that may be unacceptable

... assessment

4. another attraction of Bayesian arguments, in principle at least, is the assimilation of external evidence
 - ▶ most applications of objective approaches use some form of reference prior representing vague knowledge
 - ▶ this is increasingly questionable as the dimension of the parameter space increases
 - ▶ important questions concerning higher order matching as the number of parameters increases seem open
5. a view that does not accommodate some form of model checking, even if very informally, is inadequate

Classical Principles

- ▶ **sufficiency**: if $f(y; \theta) \propto f_1(s; \theta)f_2(t | s)$, inference about θ should be based on s ✓
- ▶ **ancillary**: if $f(y; \theta) \propto f_1(s | t; \theta)f_2(t)$, inference about θ should be based on s , given t ✓?
 - ▶ relevant subsets
- ▶ **likelihood**: inference should be based on the likelihood function ×
 - likelihood as equivalence class of functions of θ
- ▶ × for direct use of the likelihood function, but
- ▶ inference constructed from the likelihood function seems to be widely accepted

... classical principles

- ▶ sufficiency and ancillary become more difficult for inference about parameters of interest, in the presence of nuisance parameters
- ▶ asymptotic theory provides a means to incorporate these notions in an approximate sense
- ▶ but the details continue to be somewhat cumbersome
- ▶ Bayesian methods, being based on the observed data, can avoid this consideration
- ▶ at the expense of specification of prior probabilities
- ▶ which is **much** more difficult in the nuisance parameter setting

... classical principles

- ▶ a pivotal quantity is a function of the data, y , and parameter of interest ψ , with a **known** distribution
- ▶ inversion of a pivotal quantity, using its known distribution, gives a p -value function of ψ , or a confidence distribution for ψ ; i.e. a set of **nested** confidence regions at any desired confidence level
- ▶ the standardized maximum likelihood estimator is an example of such a pivotal quantity
- ▶

$$\begin{aligned}(\hat{\theta} - \theta)/\hat{\sigma}_{\theta} &\sim N(0, 1) \\ 2\{\ell(\hat{\theta}) - \ell(\theta)\} &\sim \chi_d^2\end{aligned}$$

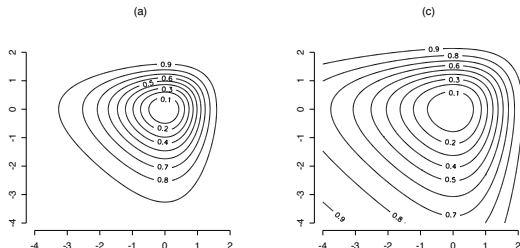
$$\hat{\sigma}_{\theta}^2 = -\ell''(\hat{\theta})^{-1}$$

Some insight from asymptotic theory

- ▶ we can construct pivotal quantities to a higher order of approximation
- ▶ for example,

$$2\{\ell(\hat{\theta}) - \ell(\theta)\} / \{1 + B(\theta)/n\} \sim \chi_d^2 \quad O(n^{-2})$$

Bartlett 1937



... asymptotic theory

- ▶ we can do better, if the parameter of interest ψ is scalar; nuisance parameters λ vector



$$r(\psi) = [2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\}]^{1/2} \sim N(0, 1)$$

$$O(n^{-1/2})$$



$$r^*(\psi) = r(\psi) + \frac{1}{r(\psi)} \log \left\{ \frac{Q(\psi)}{r(\psi)} \right\} \sim N(0, 1)$$

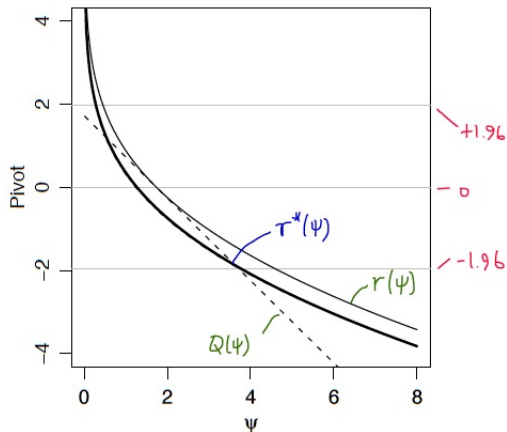
$$O(n^{-3/2})$$

- ▶ a large deviation result; leads to very accurate inferences

Brazzale et al. 2008

$$\theta = (\psi, \lambda) \quad \ell_p(\psi) = \ell(\psi, \hat{\lambda}_\psi) \quad Q(\psi) = r(\psi) + o_p(1)$$

... asymptotic theory



... asymptotic theory

- ▶ leads to insight about the point of departure between Bayesian and frequentist methods

Pierce & Peters, 1994; Fraser et al 2010

- ▶ because, the corresponding Bayesian pivot is

$$r^*(\psi) = r(\psi) + \frac{1}{r(\psi)} \log \left\{ \frac{Q^\pi(\psi)}{r(\psi)} \right\}$$

- ▶ any prior $\pi(\theta)$ for which $Q(\psi) = Q_B^\pi(\psi)$ ensures Bayesian inference calibrated for ψ

... asymptotic theory

- ▶ leads to insight about the point of departure between Bayesian and frequentist methods

Pierce & Peters, 1994; Fraser et al 2010

- ▶ because, the corresponding Bayesian pivot is

$$r^*(\psi) = r(\psi) + \frac{1}{r(\psi)} \log \left\{ \frac{Q^\pi(\psi)}{r(\psi)} \right\}$$

- ▶ any prior which is calibrated for ψ is not likely to be calibrated for other components of θ

... asymptotic theory

- ▶ leads to insight about the point of departure between Bayesian and frequentist methods

Pierce & Peters, 1994; Fraser et al 2010

- ▶ because, the corresponding Bayesian pivot is

$$r^*(\psi) = r(\psi) + \frac{1}{r(\psi)} \log \left\{ \frac{Q^\pi(\psi)}{r(\psi)} \right\}$$

- ▶ adjustments for high-dimensional nuisance parameters are the most important ingredient
- ▶ these adjustments are built into $Q(\psi)$

... asymptotic theory

- ▶ leads to insight about the point of departure between Bayesian and frequentist methods

Pierce & Peters 1994; Fraser et al 2010

- ▶ because, the corresponding Bayesian pivot is

$$r^*(\psi) = r(\psi) + \frac{1}{r(\psi)} \log \left\{ \frac{Q^\pi(\psi)}{r(\psi)} \right\}$$

- ▶ can be replicated by bootstrap sampling under $(\psi, \hat{\lambda}_\psi)$

DiCiccio & Young 2008 ; Fraser & Rousseau 2008

Foundations and Applications


- ▶ “Foundations must be continually tested against (new) applications”
- ▶ in spite of the likelihood principle, a great deal of applied work considers the distribution of quantities based on the likelihood function (maximum likelihood estimator, likelihood ratio statistic, etc.)
- ▶ in spite of theorems on coherence and exchangeability, a great deal of applied work with Bayesian methods uses what are hoped to be “non-influential” priors
- ▶ the question is whether or not there really are non-influential
- ▶ “non-informative priors are the perpetual motion machine of statistics”

Wasserman 2012

... foundations and applications

- ▶ Are we ready for “Big Data”
- ▶ Will statistical principles be helpful?
- ▶ Are the classical principles enough?
- ▶ “Inferential giants”: assessment of sampling bias, inference about tails, resampling inference, change point detection, reproducibility of analyses, causal inference for observational data, efficient inference for temporal streams






This PDF is available from The National Academies Press at http://www.nap.edu/catalog.php?record_id=18374



Frontiers in Massive Data Analysis

ISBN
978-0-309-28776-4
129 pages
6 x 9
PAPERBACK (2013)

Committee on the Analysis of Massive Data; Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences and Their Applications; Division on Engineering and Physical Sciences; National Research Council


 Add book to cart  Find similar titles  Share this PDF    

National Academies Press http://www.nap.edu/catalog.php?record_id=18374

... foundations and applications

- ▶ Are we ready for “Big Data”
- ▶ Will statistical principles be helpful?
- ▶ Are the classical principles enough?
- ▶ “Inferential giants”: assessment of sampling bias, **inference about tails, resampling inference, change point detection, reproducibility of analyses, causal inference for observational data**, efficient inference for temporal streams






This PDF is available from The National Academies Press at http://www.nap.edu/catalog.php?record_id=18374



Frontiers in Massive Data Analysis

ISBN
978-0-309-28776-4
129 pages
6 x 9
PAPERBACK (2013)

Committee on the Analysis of Massive Data; Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences and Their Applications; Division on Engineering and Physical Sciences; National Research Council

 Add book to cart  Find similar titles  Share this PDF    

National Academies Press http://www.nap.edu/catalog.php?record_id=18374

Conclusion

- ▶ **from our abstract:** “statistical theory serves to provide a systematic way of approaching new problems, and to give a common language for summarizing results”
- ▶ “ideally this foundation and common language ensures that the statistical aspects of one study or of several studies on closely related phenomena can, in broad terms, be readily understood by the non-specialist”
- ▶ “However, the continuing distinction between ‘Bayesians’ and ‘frequentists’ is a source of ambiguity and potential confusion”
- ▶ this does not bely the utility of methods that combine probabilities using Bayes’ theorem
- ▶ ‘Math Stat’, the course students love to hate, is more important than ever!