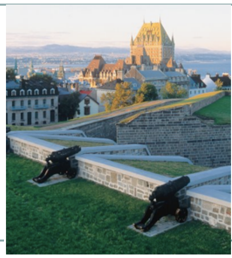
Thoughts on the theory of statistics

Nancy Reid

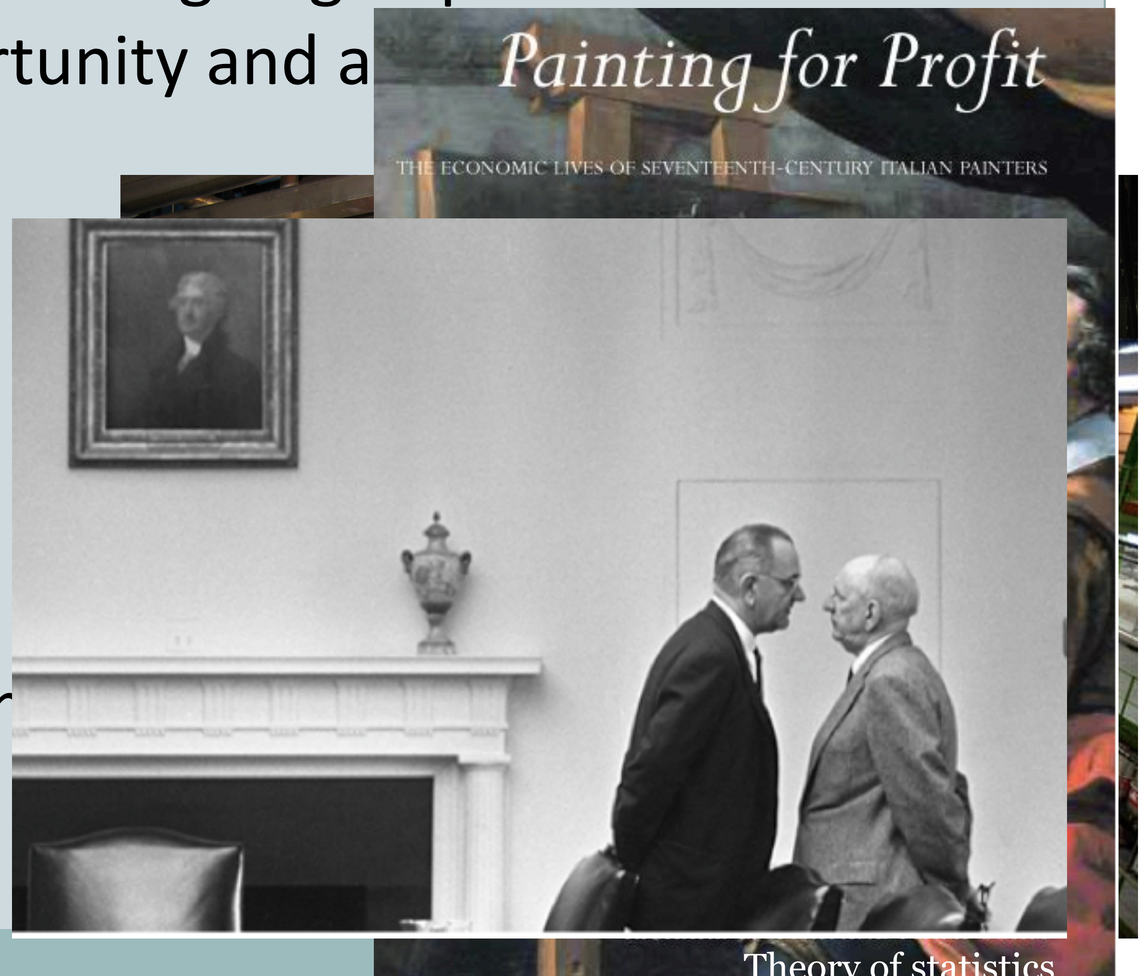
2010 Annual Meeting in Québec City

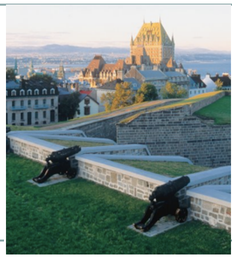
38th Annual Meeting of the Statistical Society of Canada



Statistics in demand

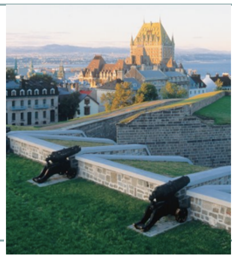
- “Statistical science is undergoing unprecedented growth in both opportunity and a
- High energy physics
- Art history
- Reality mining
- Bioinformatics
- Complex surveys
- Climate and environm
- SSC 2010 …





Statistical Thinking

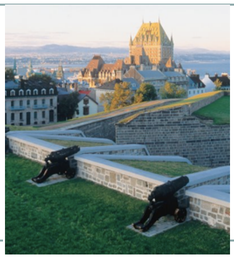
- Dramatic increase in resources now available



Statistical Thinking

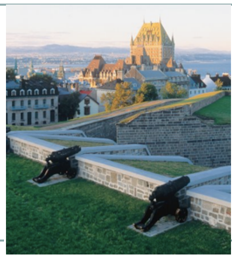
- If a statistic was the answer, what was the question?

- Percentages and risk
  - relative and absolute change



Statistical theory for 20xx

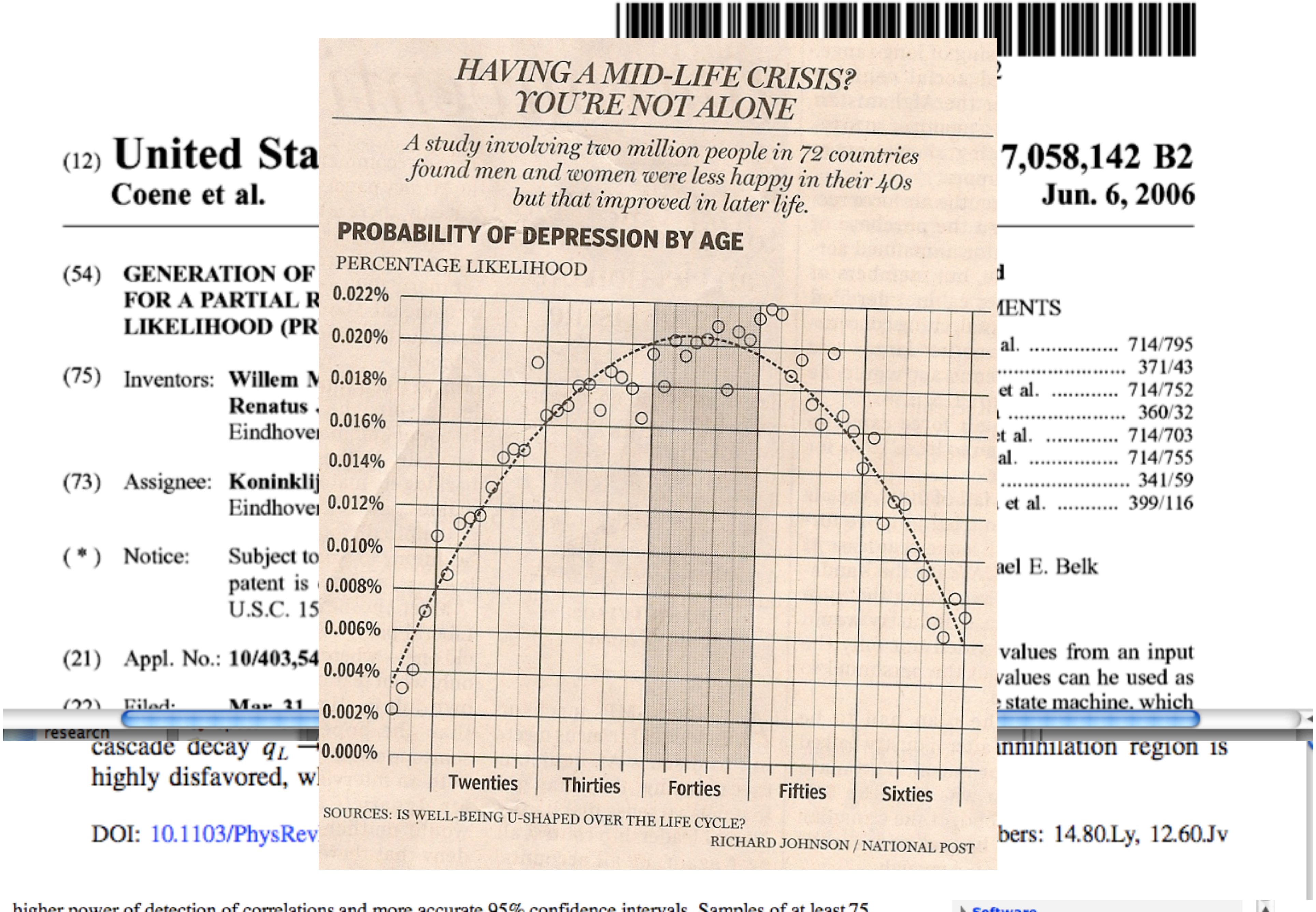
- What should we be teaching?
- If a statistic was the answer, what was the question?
  - Design of experiments and surveys
- Common pitfalls
  - Summary statistics: sufficiency etc.
- How sure are we?
  - Inference
- Percentages and risk
  - Interpretation

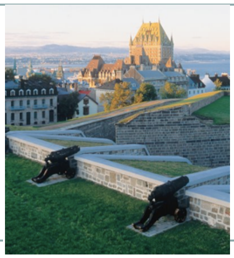


Models and likelihood

- Modelling is difficult and important
- We can get a lot from the likelihood function
- Not only point estimators $\hat{\theta}$
- Not only (not at all!!) most powerful tests $\dfrac{f(y;\theta_1)}{f(y;\theta_0)}$
- Inferential quantities (pivots)
- Inferential distributions (asymptotics)
- A natural starting point, even for very complex models

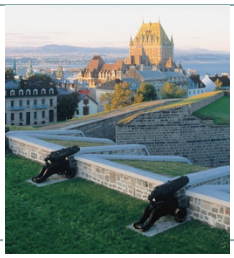
Likelihood is everywhere!



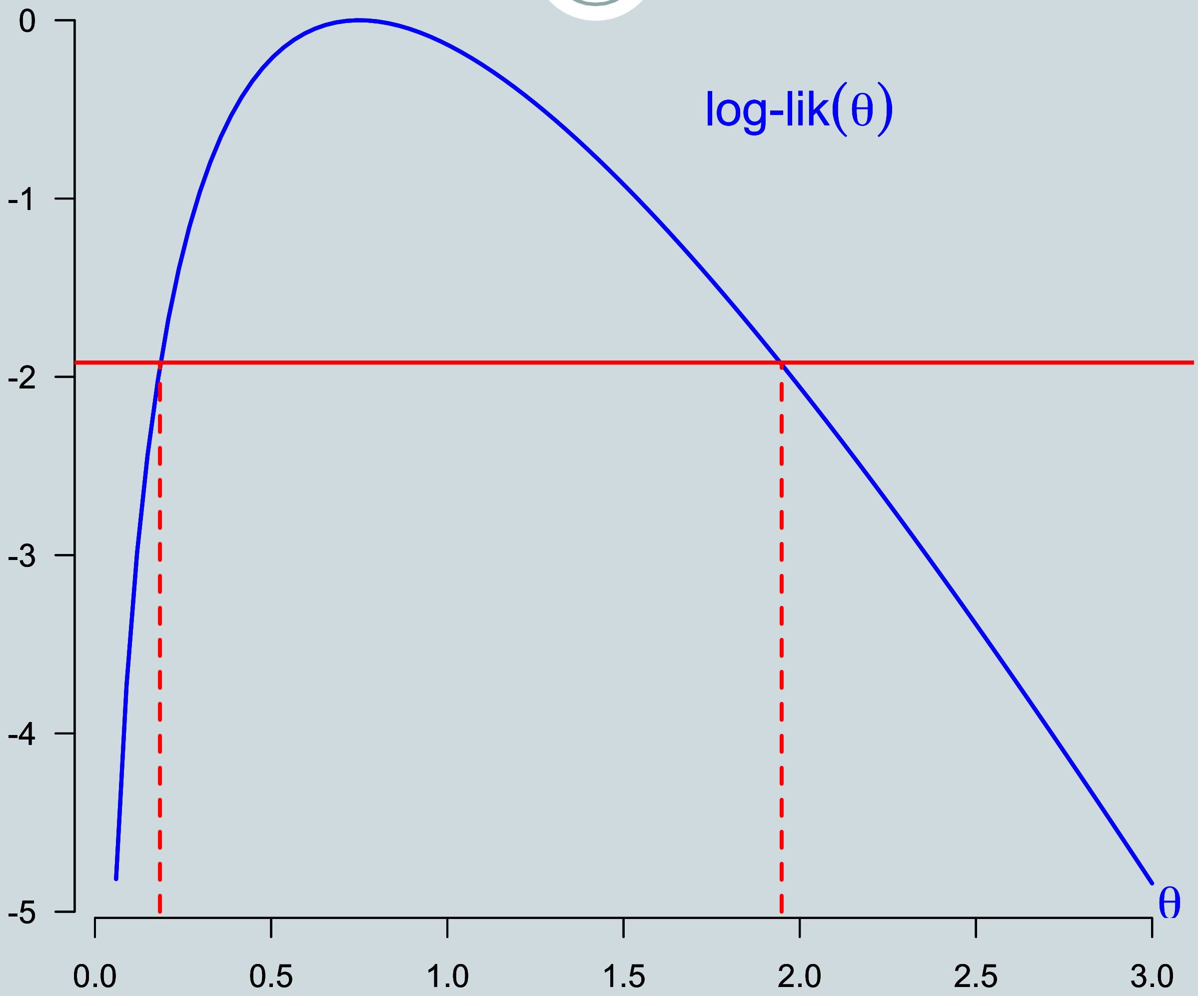


Outline

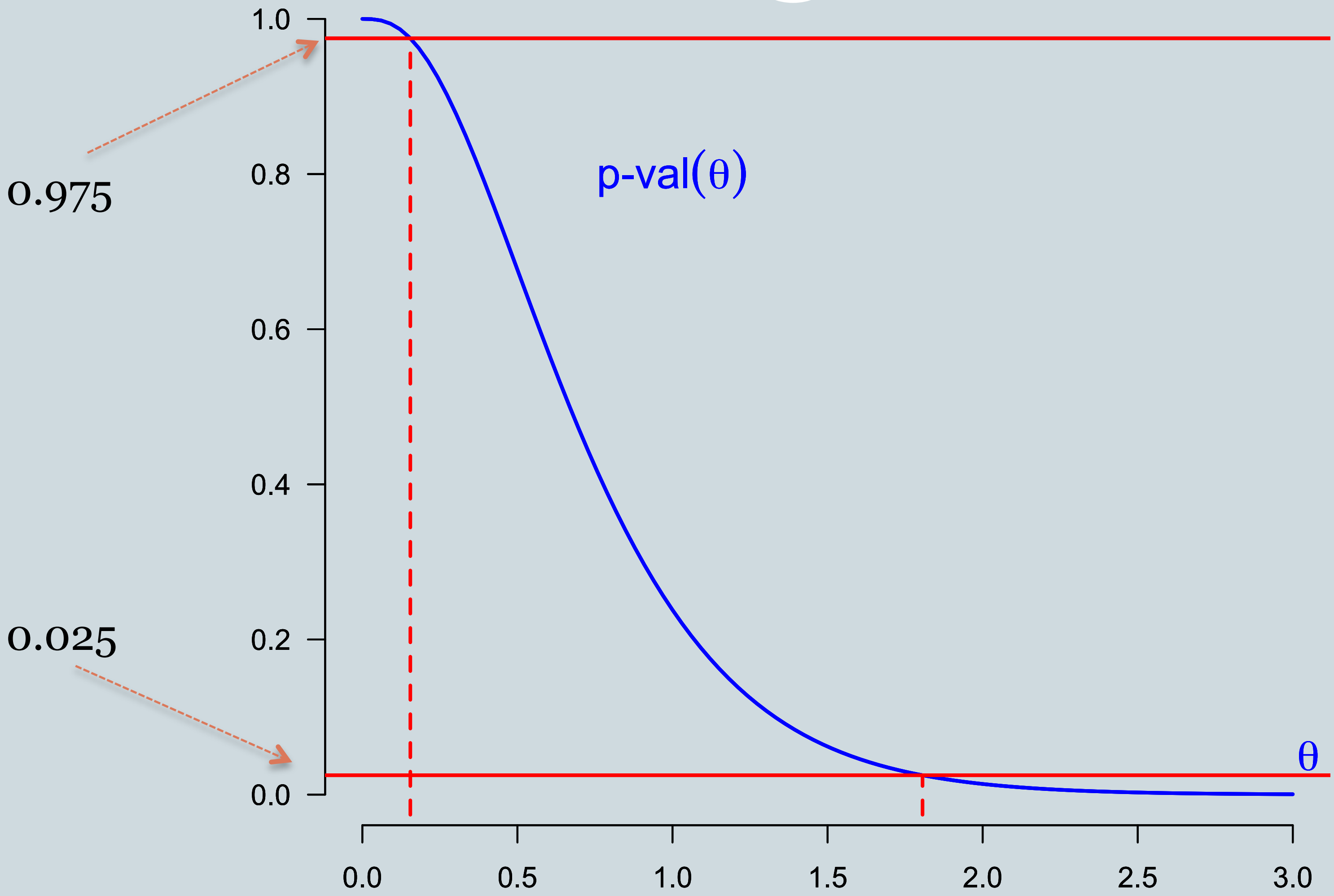
1. Higher order asymptotics
   likelihood as pivotal
2. Bayesian and non-Bayesian inference
3. Partial, quasi, composite likelihood
4. Where are we headed?

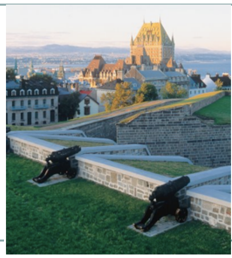


P-value functions from likelihood



P-value functions from likelihood



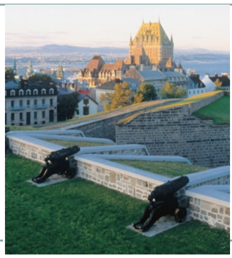


Can be nearly exact

- Likelihood root $r(\theta) = \pm\sqrt{[2\{\ell(\hat\theta) - \ell(\theta)\}]}$
- Maximum likelihood estimate $q(\theta) = (\hat\theta - \theta) j^{1/2}(\hat\theta)$
- Score function $s(\theta) = \ell'(\theta) j^{-1/2}(\hat\theta)$
- All approximately distributed as $N(0,1)$

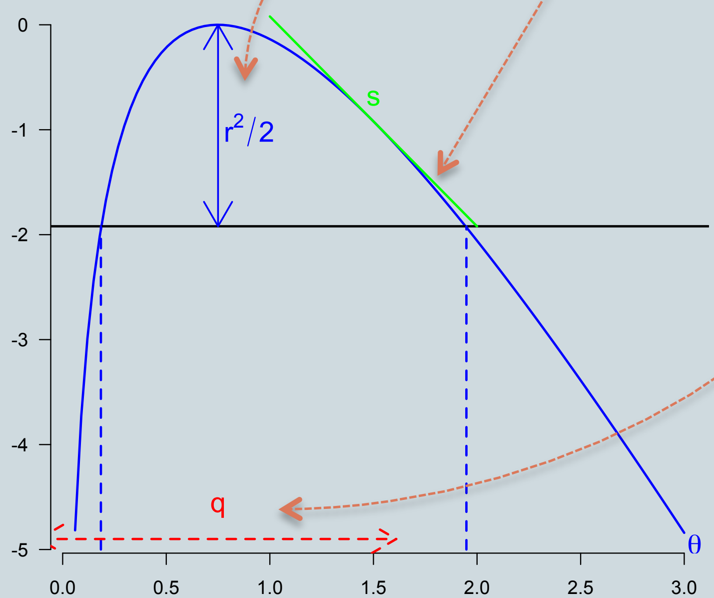
Much better : $r^*(\theta) = r(\theta) + \dfrac{1}{r(\theta)} \log \dfrac{Q(\theta)}{r(\theta)}$

- $Q(\theta)$ can be $q(\theta)$ or $s(\theta)$ or ...

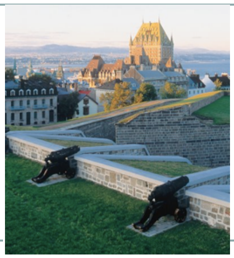


Can be nearly exact

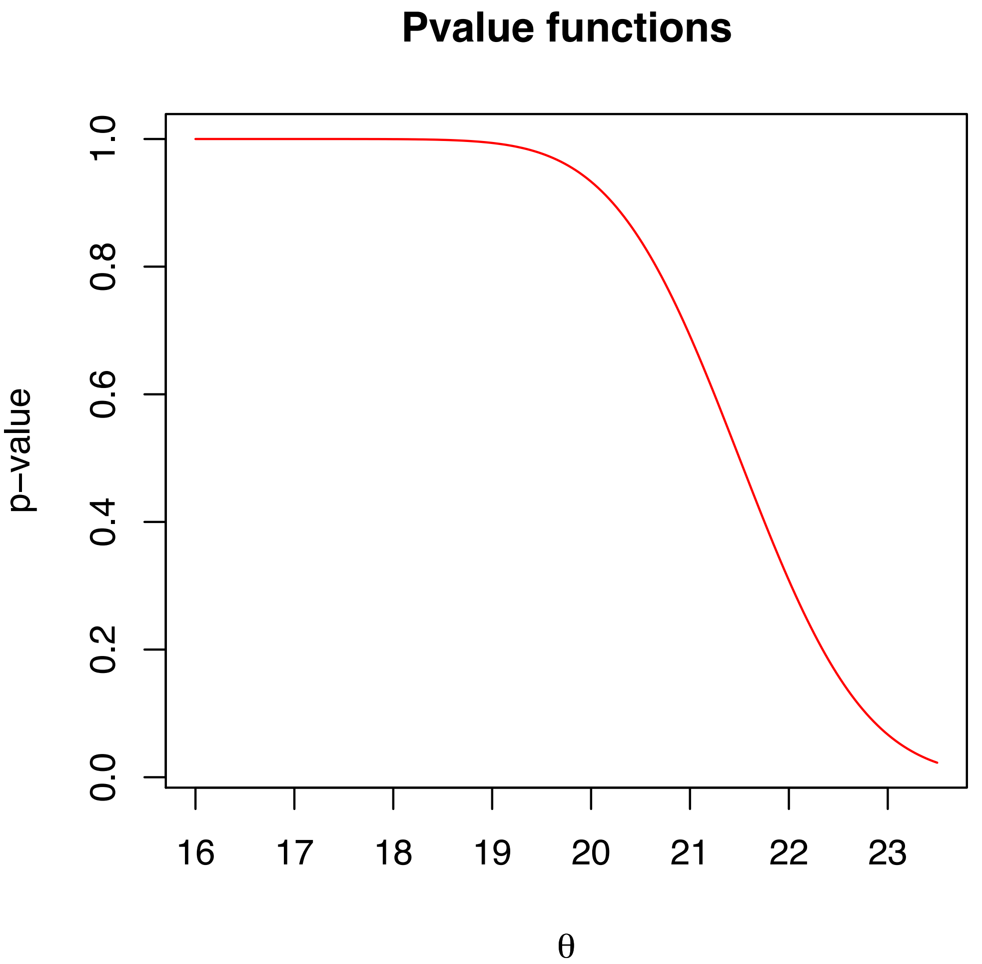
- Likelihood root $r(\theta) = \pm\sqrt{[2\{\ell(\hat\theta) - \ell(\theta)\}]}$
- Maximum likelihood estimate $q(\theta) = (\hat\theta - \theta) j^{1/2}(\hat\theta)$
- Score function $s(\theta) = \ell'(\theta) j^{-1/2}(\hat\theta)$

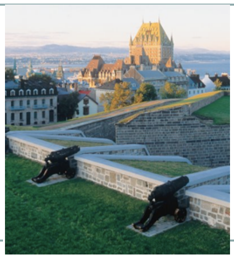


$$j(\theta) = -\ell''(\theta)$$

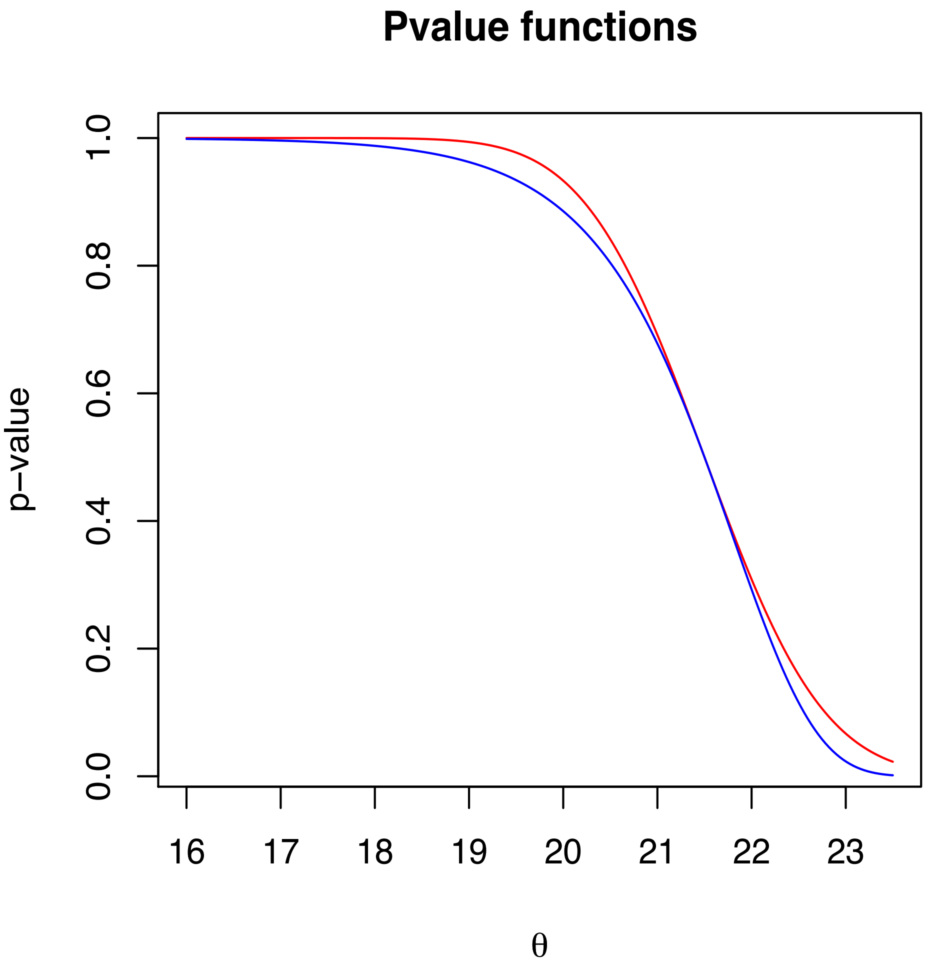


Can be nearly exact

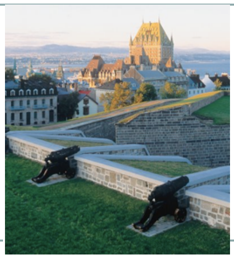


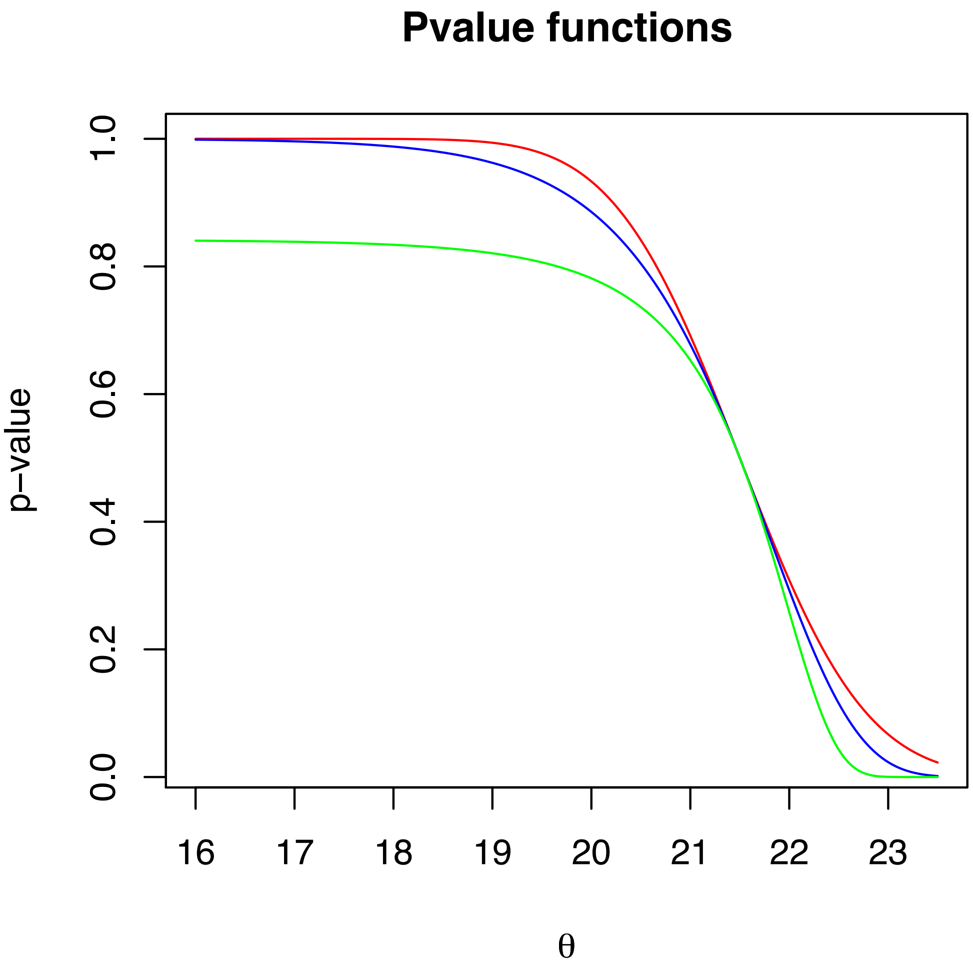


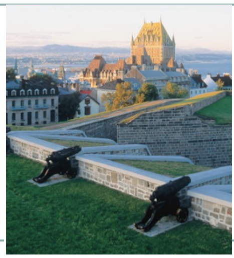
Can be nearly exact



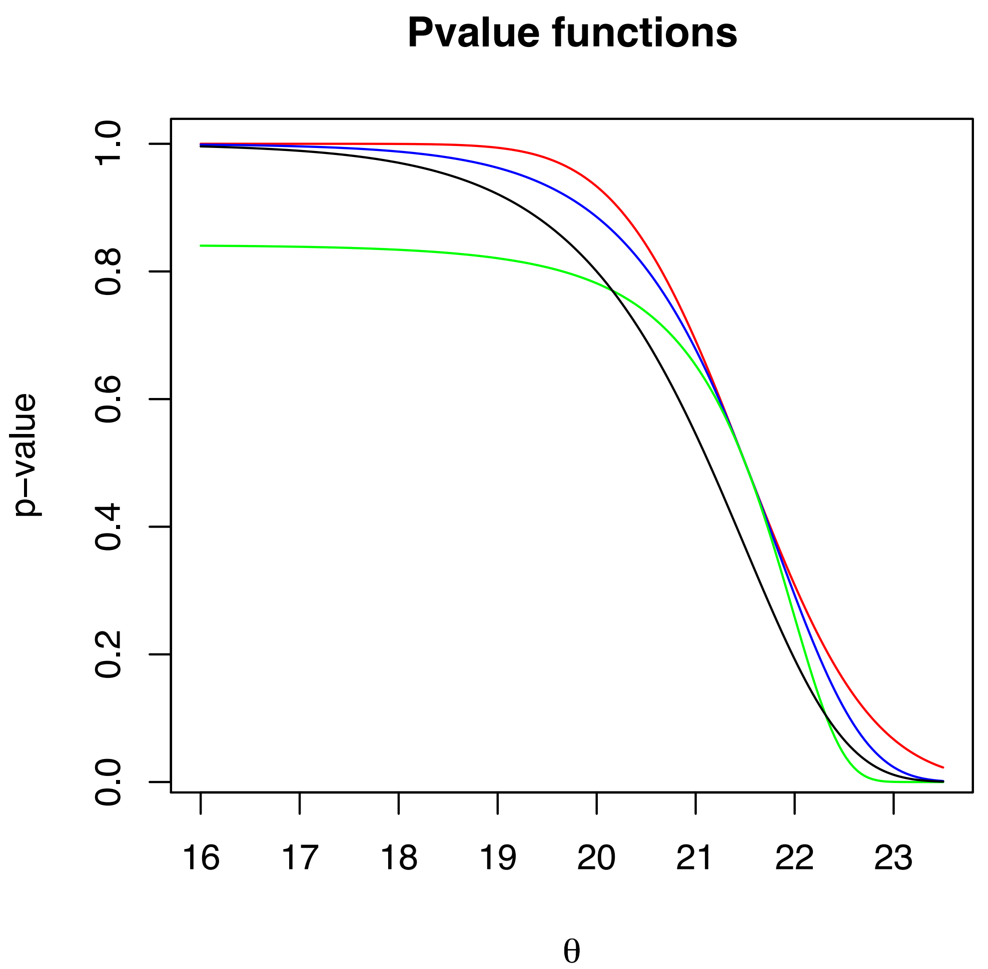
Can be nearly exact

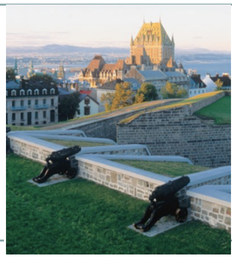






Can be nearly exact

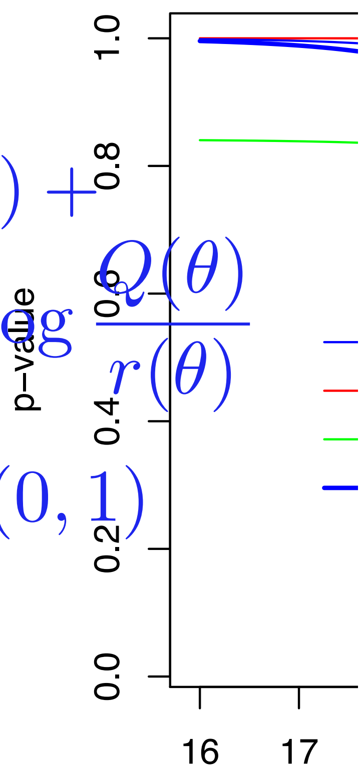


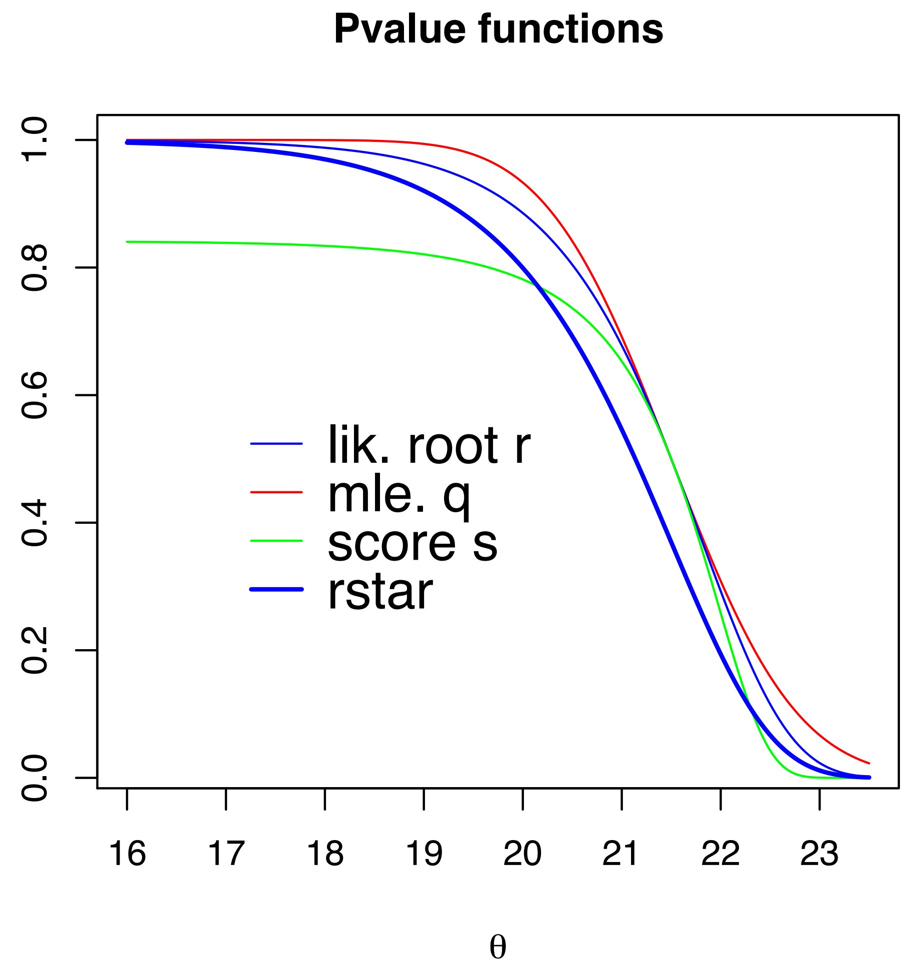


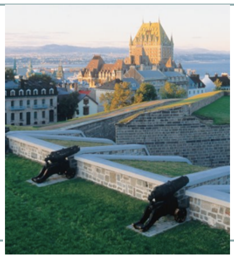
Can be nearly exact

$$r^*(\theta) = r(\theta) + \frac{1}{r(\theta)} \log \frac{Q(\theta)}{r(\theta)}$$

$$\simeq N(0,1)$$

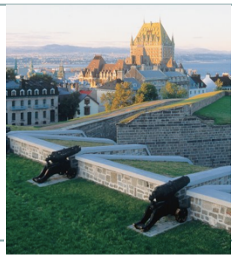




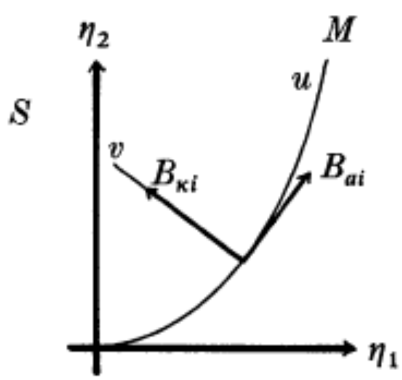


Using higher order approximations

- Excellent approximations for ‘easy’ cases
  - Exponential families, non-normal linear regression
- More work to construct for ‘moderate’ cases
  - Autoregressive models, fixed and random effects, discrete responses
- Fairly delicate for ‘difficult’ cases
  - Complex structural models with several sources of variation
- Best results for scalar parameter of interest
  - But we may need inference for vector parameters



Where does this come from?



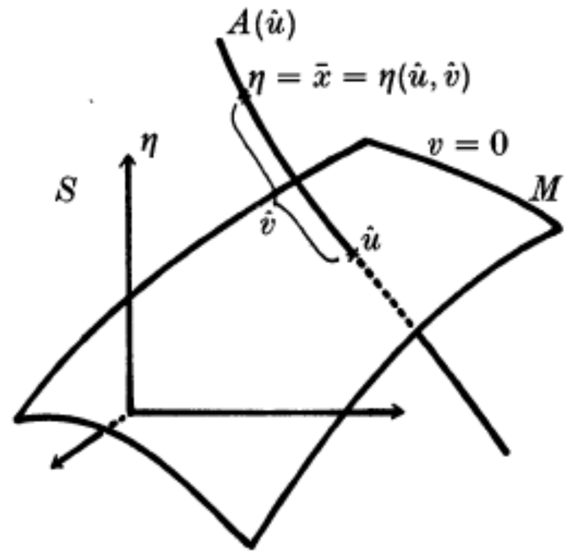
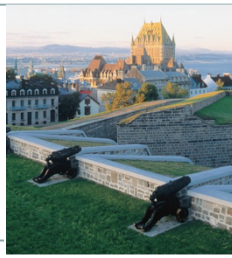


Fig. 1 (left). Example of curved exponential family $N(u, a^2 u^2)$.
Fig. 2 (right). Ancillary subspace $A(u)$ and local coordinates (u, v).

[4]Amari, 1982, Biometrika; Efron, 1975, Annals



Where does this come from?

5,6,7

- Differential geometry of statistical models
- Theory of exponential families
- Edgeworth and saddlepoint approximations
- Key idea:
- A smooth parametric model can be approximated
  by a tangent exponential family model
- Requires differentiating log-likelihood function
  on the sample space
- Permits extensions to more complex models

Where does this come from?

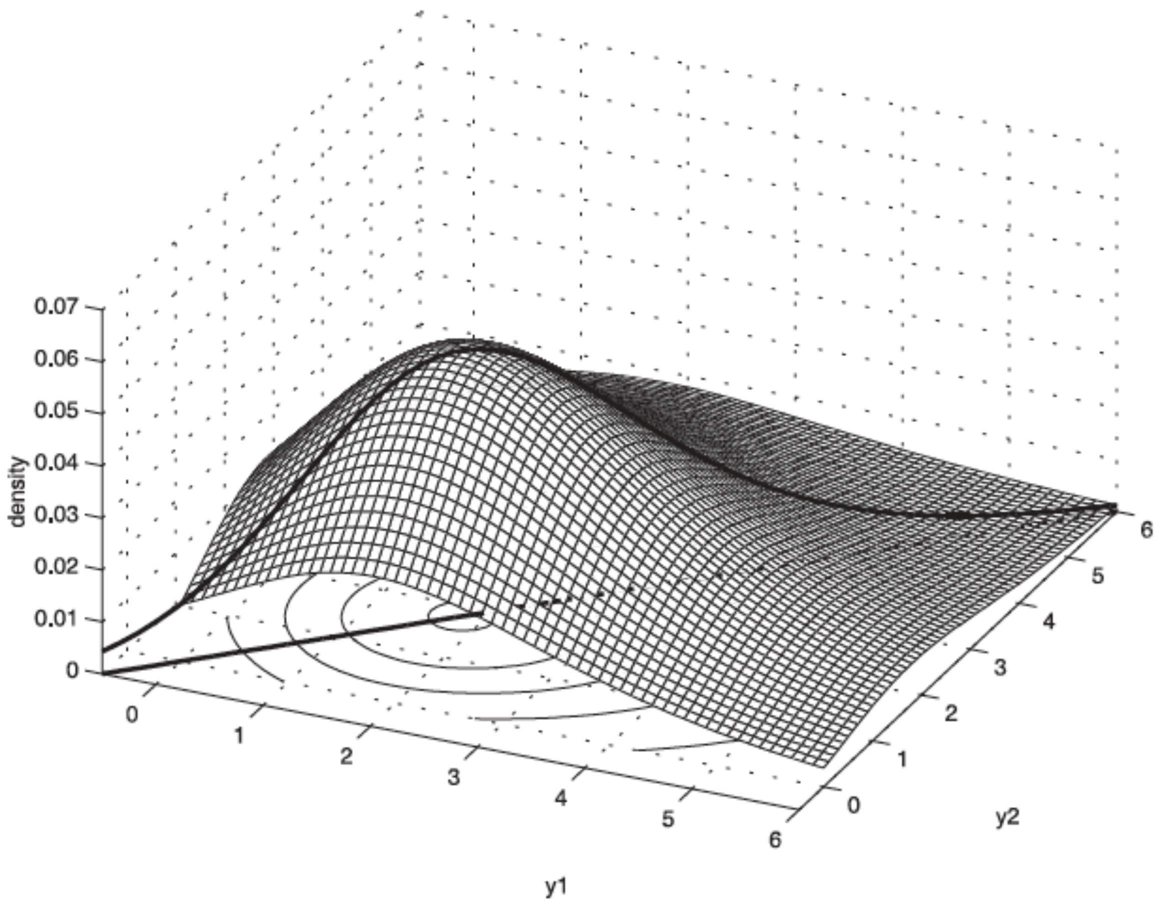
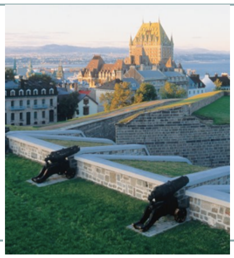


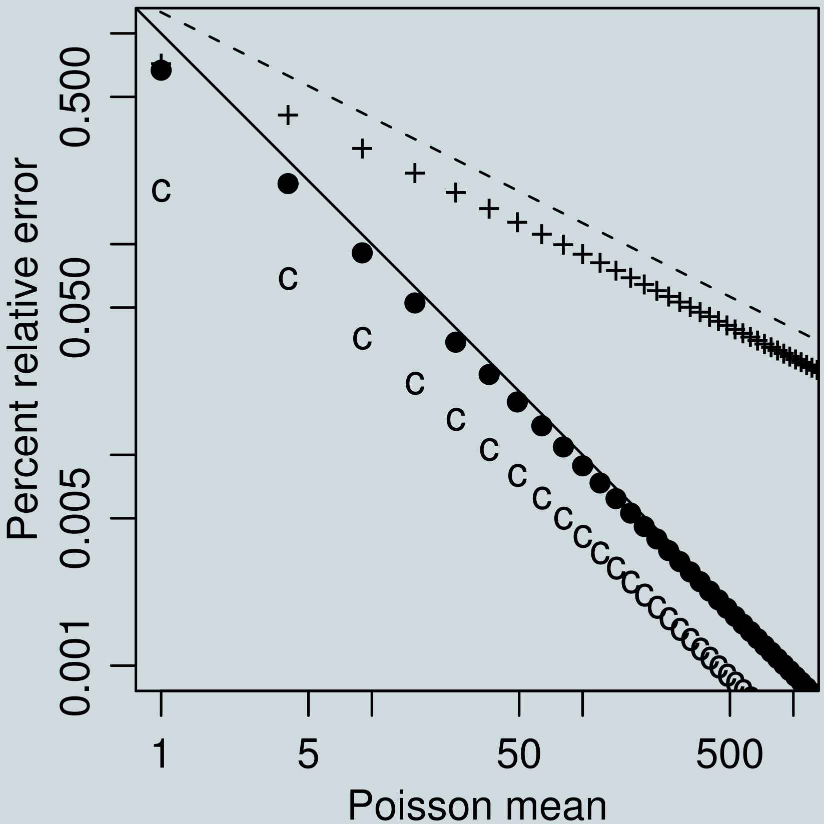
FIG. 2. *The second-order ancillary, with tangent vectors given by V, is constant along the solid curve in the (y_1, y_2) plane.*

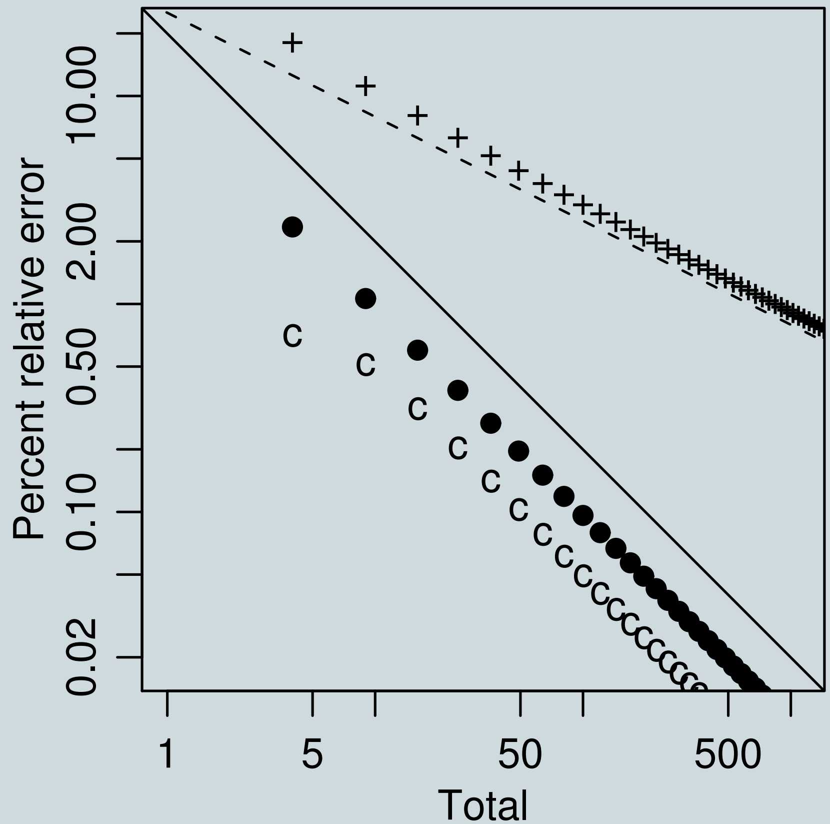


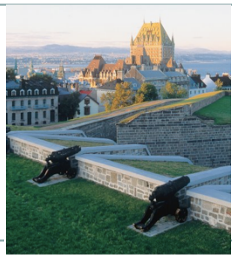
Generalizations

- To discrete data
- Where differentiating the log-likelihood on the sample space is more difficult
- Solution: use expected value of score statistic instead
- Relative error $O(n^{-1})$ instead of $O(n^{-3/2})$
- Still better than the normal approximation

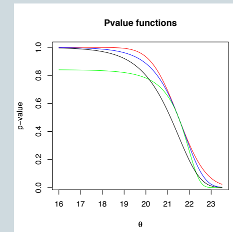
Generalizations



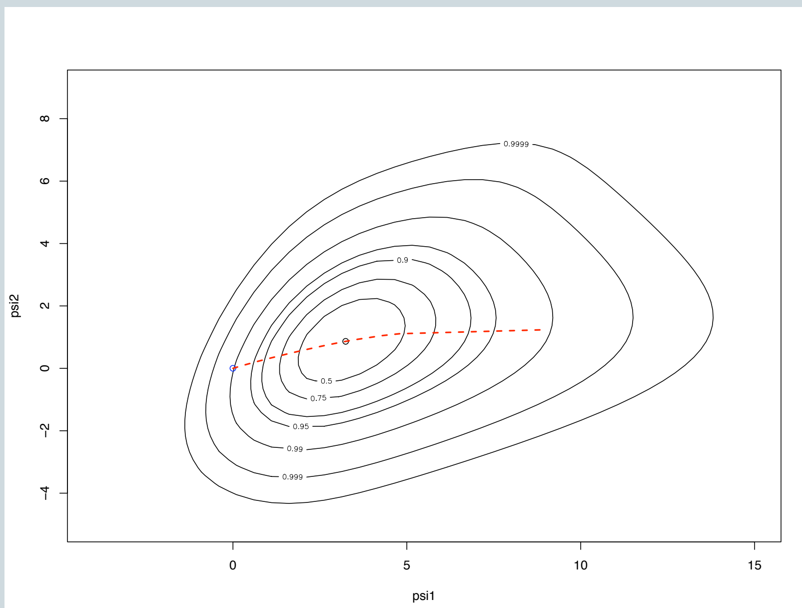


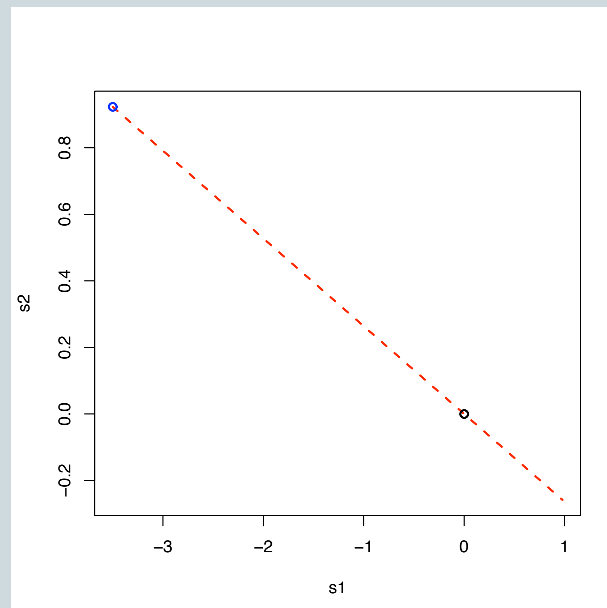


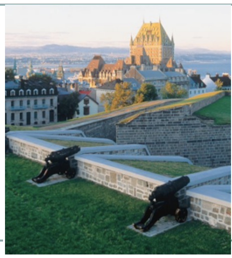
Generalizations



- To vector parameters of interest
- But our solutions require a single parameter
- Solution: use length of the vector, conditioned on the direction

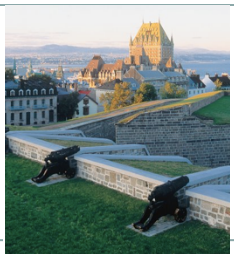






Generalizations

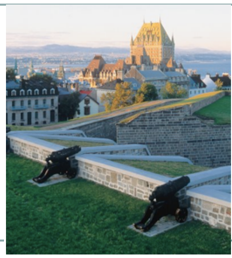
- Extending the role of the exponential family
- By generalizing differentiation on the sample space
- Idea: differentiate the expected log-likelihood
  - Instead of the log-likelihood
- Leads to a new version of approximating exponential family
- Can be used with pseudo-likelihoods



What can we learn?

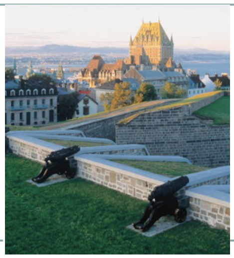
- Higher order approximation requires
- Differentiating the log-likelihood function

  on the sample space
- Bayesian inference will be different
- Asymptotic expansion highlights the discrepancy
- Bayesian posteriors are in general not calibrated
- Cannot always be corrected by choice of the prior
- We can study this by comparing Bayesian and nonBayesian approximations



Example: inference for ED50

- Logistic regression with a single covariate
- On the logistic scale $\Pr(y_i = 1) = \alpha + \beta x_i$
- Use flat priors for (α, β)
- Parameter of interest is $\psi = -\alpha/\beta$
- Empirical coverage of Bayesian posterior intervals:
  - 0.90, 0.88, 0.89, 0.90
- Empirical coverage of intervals using $\Phi(r^*)$
  - 0.95, 0.95, 0.95, 0.95



Flat priors are not a good idea!

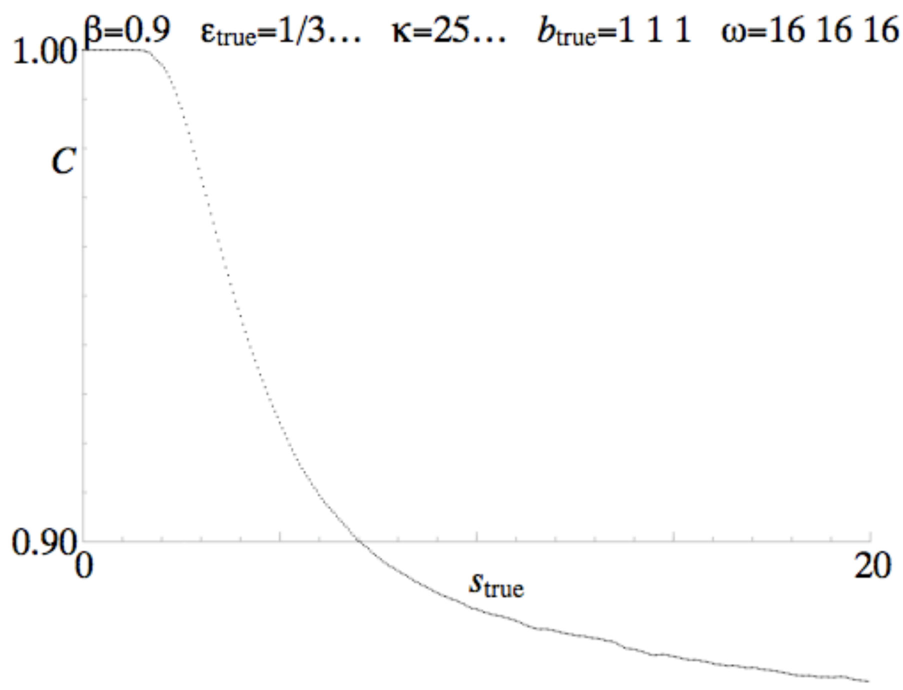
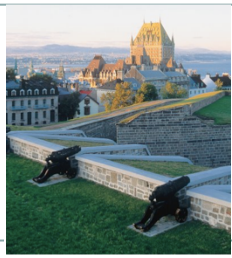
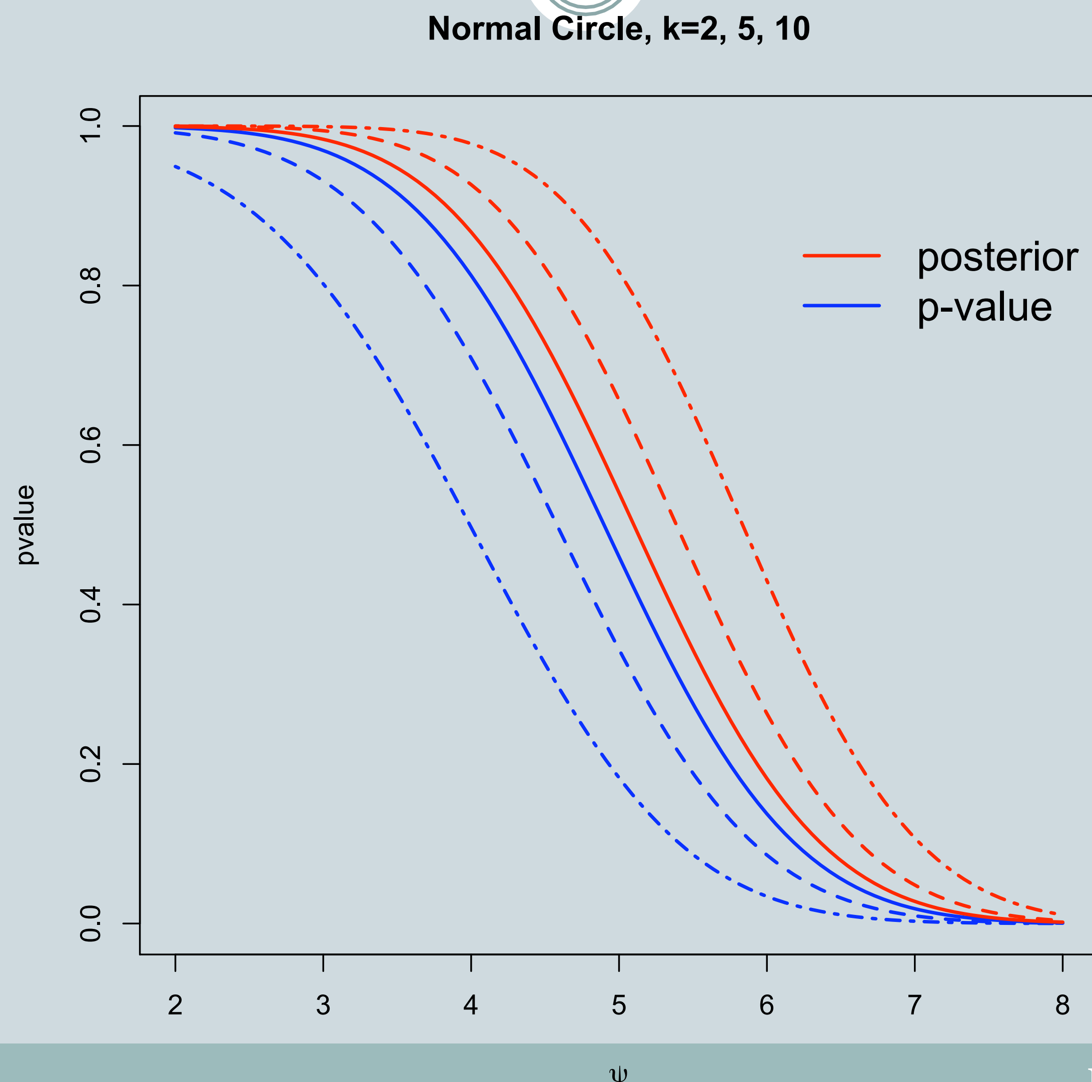
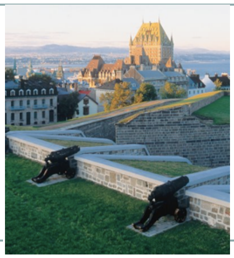


Fig. 6. 3 independent channels. Coverage for 90% credibility level upper limits, acceptance uncertainty = 34%/channel, background uncertainty = 25%/channel.



Flat priors are not a good idea!





Flat priors are not a good idea!

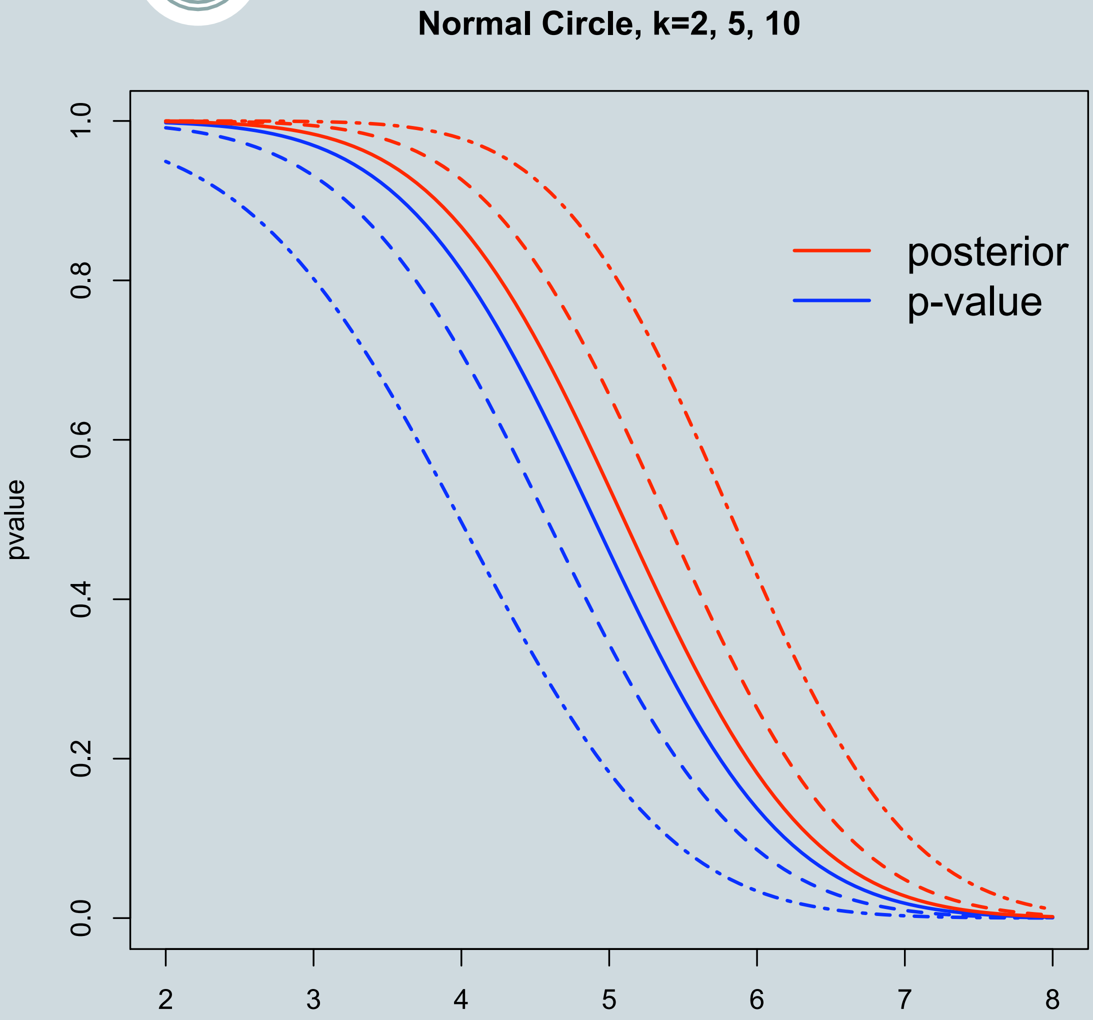
Bayesian p-value –

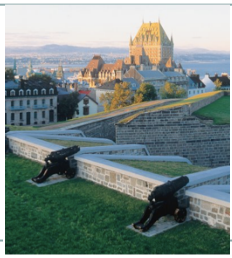
Frequentist p-value

$$\simeq \frac{(k-1)}{\psi\sqrt{n}}$$

$$Y_i \sim N(\mu_i, \frac{1}{n})$$

$$\psi = (\sum_{i=1}^{k} \mu_i^2)^{1/2}$$

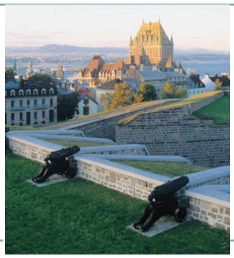




More complex models

- Likelihood inference has desirable properties
- Sufficiency, asymptotic efficiency
- Good approximations to needed distributions
- Derived naturally from parametric models
- Can be difficult to construct,

  especially in complex models
- Many natural extensions: partial likelihood for censored data, quasi-likelihood for generalized estimating equations, composite likelihood for dependent data

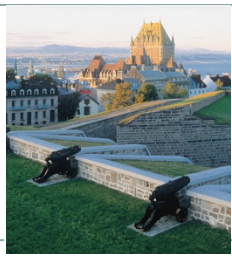


Complex models

- Example: longitudinal study of migraine sufferers
- Latent variable $Y_{ij}^* = x_{ij}^T \beta + U_i + \epsilon_{ij}$
- Observed variable

$$y_{ij} \in \{1, \ldots, h\} \leftrightarrow \alpha_{y_{ij}-1} < Y_{ij}^* < \alpha_{y_{ij}}$$

- E.g. no headache, mild, moderate, intense …
- x_{ij} Covariates: age, education, painkillers, weather, …
- $U_i, \; \epsilon_{ij}$ random effects between and within subjects
- Serial correlation $\epsilon_{ij} = \rho \epsilon_{i,j-1} + (1-\rho^2)^{1/2} \eta_{ij}$



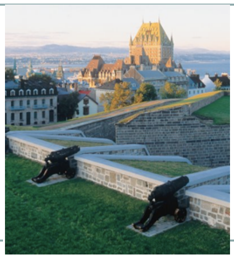
Likelihood for longitudinal discrete data

- Likelihood function

$$L(\theta; y) = \prod_{i=1}^{n} \int \cdots \int \phi_{m_i}(z_{i1}, \ldots, z_{im_i}; R)\mathrm{d}z_{i1} \ldots \mathrm{d}z_{im_i}$$

- Hard to compute
- Makes strong assumptions
- Proposal: use bivariate marginal densities

  instead of full multivariate normal densities
- Giving a mis-specified model

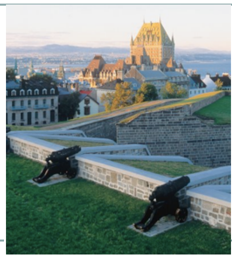


Composite likelihood

- Composite likelihood function

$$CL(\theta; y) = \prod_{i=1}^{n} \prod_{j<k} \int\int \phi_2(z_{i1}, z_{i2}; R_2) \mathrm{d}z_{i1} \mathrm{d}z_{i2}$$

- More generally $CL(\theta) = \prod_{i=1}^{n} \prod_{k=1}^{K} f(y_i \in \mathcal{A}_k)$
- Sets $\mathcal{A}_k$ index marginal or conditional (or …) distributions
- Inference based on theory of estimating equations

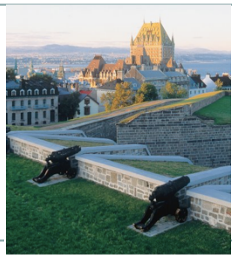


A simple example

$$Y_i \sim N_p(0, \Sigma)$$
$$i = 1, \ldots, n$$

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \theta & \ldots & \theta \\ \theta & 1 & \ldots & 1 \\ & & \vdots & \\ \theta & \theta & \ldots & 1 \end{pmatrix}$$

- Pairwise likelihood estimator of θ fully efficient
- If $\sigma^2 = 1$, loss of efficiency depends on dimension p
- Small for dimension less than, say, 10
- Falls apart if $p \rightarrow \infty$ for fixed sample size
  - Relevant for time series, genetics applications



Composite likelihood estimator

$$CL(\theta) = \prod_i \prod_k f(y_i \in \mathcal{A}_k; \theta)$$

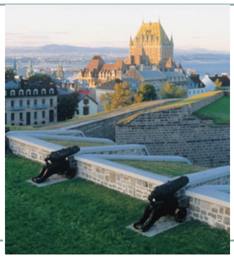
$$\hat{\theta}_{CL} \xrightarrow{p} \infty$$

$$\sqrt{n}(\hat{\theta}_{CL} - \theta) \xrightarrow{d} N\{0, G^{-1}(\theta)\}$$

$$G(\theta) = J(\theta)H^{-1}(\theta)J(\theta)$$

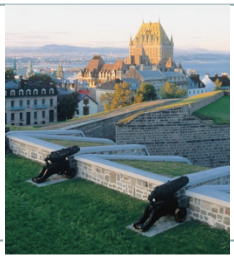
Godambe information

$$J(\theta) = E\{-\partial^2 CL(\theta)/\partial\theta^2\}, \quad H(\theta) = E\{\partial CL(\theta)/\partial\theta\}^2$$



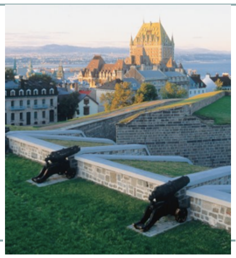
Recent Applications

- Longitudinal data, binary and continuous: random effects models
- Survival analysis: frailty models, copulas
- Multi-type responses: discrete and continuous; markers and event times
- Finance: time-varying covariance models
- Genetics/bioinformatics: CCL for vonMises distribution: protein folding; gene mapping; linkage disequilibrium
- Spatial data: geostatistics, spatial point processes

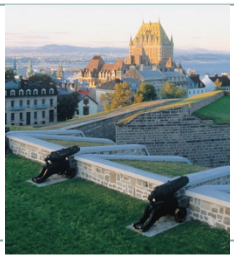


… and more

- Image analysis
- Rasch model
- Bradley-Terry model
- State space models
- Population dynamics
- …

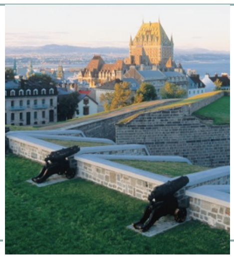


What can we learn?



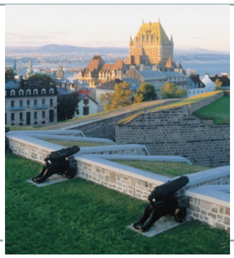
What do we need to know?

- Why are composite likelihood estimators efficient?
- How much information should we use?
- Are the parameters guaranteed to be identifiable?
- Are we sure the components are consistent with a ‘true’ model?
- Can we make progress if not?
- How do joint densities get constructed?
- What properties do these constructions have?
- Is composite likelihood robust?



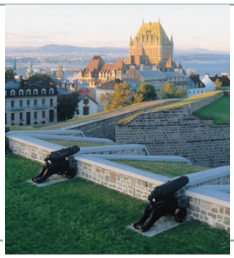
Why is this important?

- Composite likelihood ideas generated from applications
- Likelihood methods seem too complicated
- A range of application areas all use the same/similar ideas
- Abstraction provided by theory allows us to step back from the particular application
- Get some understanding about when the methods might not work
- As well as when they are expected to work well



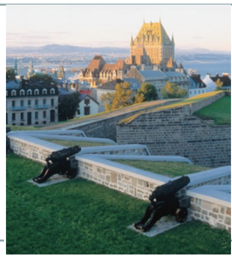
The role of theory

- Abstracts the main ideas
- Simplifies the details
- Isolates particular features
- In the best scenario, gives new insight into what underlies our intuition
- Example: curvature and Bayesian inference
- Example: composite likelihood
- Example: false discovery rates



False discovery rates

- Problem of multiple comparisons
  - Simultaneous statistical inference – R.G. Miller, 1966
- Bonferroni correction too strong
- Benjamini and Hochberg, 1995
- Introduce False Discovery Rate
  - An improvement (huge!) on "Type I and Type II error"
- Then comes data, in this case from astrophysics
- Genovese & Wasserman collaborating with Miller and Nichol



False discovery rates

Acoustic Oscillations in the Early Universe and Today

Christopher J. Miller,[1] Robert C. Nichol,[1] David J. Batuski[2]

During its first ≃100,000 years, the universe was a fully ionized plasma with a tight coupling by Thompson scattering between the photons and matter. The trade-off between gravitational collapse and photon pressure causes acoustic oscillations in this primordial fluid. These oscillations will leave predictable imprints in the spectra of the cosmic microwave background and the present-day matter-density distribution. Recently, the BOOMERANG and MAXIMA teams announced the detection of these acoustic oscillations in the cosmic microwave background (observed at redshift ≃ 1000). Here, we compare these CMB detections with the corresponding acoustic oscillations in the matter-density power spectrum (observed at redshift ≃ 0.1). These consistent results, from two different cosmological epochs, provide further support for our standard Hot Big Bang model of the universe.

The standard model of cosmology is the Inflationary Hot Big Bang scenario. A key aspect of this model is the ease with which it explains some critical observational facts about the universe. For example, the existence of the cosmic microwave background (CMB) radiation that fills all space is simply the radio remnant of a hot early phase of the universe, i.e., when it was only ≃100,000 years old. The model also provides a natural explanation for Hubble's famous expansion, large-scale coherent structures in the mass distribution (caused by quantum effects in the early universe), as well as producing a flat global geometry for the universe (*1*). In this scenario, the distribution of matter on the largest scales is connected, through well-established physics, to the temperature fluctuations in the CMB. Thus, any independent agreement between the CMB (at redshift ≃ 1000) and the matter-density distribution (at redshift ≃ 0.1) is naturally explained by the Hot Big Bang Inflationary model.

The early universe was a plasma made up of photons, electrons, and protons, along with the so-called Dark Matter. During this period, the gravitational force from potential wells (created as a result of local curvature pertubations or dark matter clumps) causes compressions in this fluid. As the plasma collapses inward, it meets resistance from photon pressure, reversing the plasma direction and causing a subsequent rarefaction. This cycle of compression and rarefaction results in acoustic oscillations, where baryons act as a source of inertia. Compression (rarefaction) of the plasma creates hot (cold) spots in the temperature of the plasma. Because the photons and baryons are coupled through Thompson scattering, the matter-density power spectrum will also exhibit these oscillations. As the universe cooled and the photons and matter decoupled, the acoustic oscillations became frozen as oscillatory features in both the temperature and matter-density power spectra. These acoustic oscillations are a general prediction from gravitational instability models of structure formation (*2*, *3*).

The recent results from the MAXIMA and BOOMERANG CMB balloon experiments provide evidence for the first two acoustic peaks (*4*–*8*). These acoustic oscillations are the peaks and valleys in Fig. 1A. The location and amplitude of the first peak indicate that

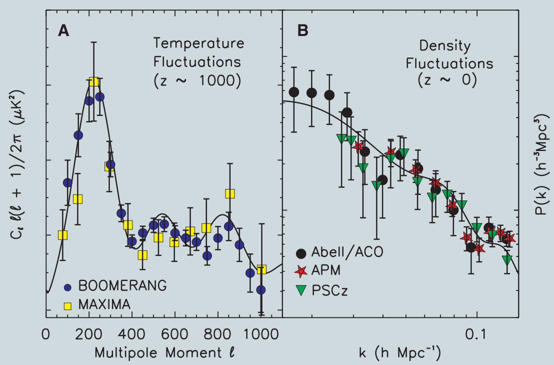
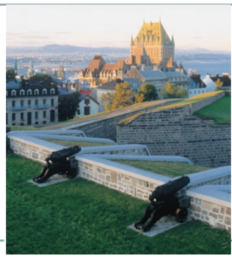


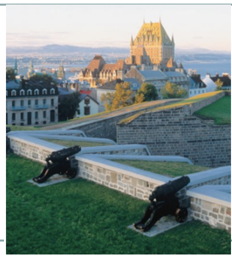
Fig. 1. We plot the CMB data from the MAXIMA and BOOMERANG experiments (**A**) alongside the matter-density data (**B**). The solid line is the best fit model ($\Omega_{matter} = 0.24$, $\Omega_{baryons} = 0.06$, and $n_s = 1.08$ with $H_0 = 69$) using the matter-density data alone. The amplitudes in both plots remain a free parameter. The solid line in (A) is not a fit to the CMB data (although the χ^2 is 34 for 32 data points). It is the resultant cosmological model using the best fit parameters from (B) and $\Omega_{vacuum} = 0.8$, consistent with the Type Ia supernovae results (*18*).

[1]Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [2]Department of Physics and Astronomy, University of Maine, Orono, ME 04469, USA.



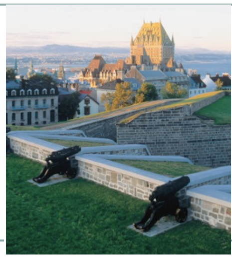
Speculation

- Composite likelihood as a smoother
- Calibration of posterior inference
- Extension of higher order asymptotics to composite likelihood
- Exponential families and empirical likelihood
- Semi-parametric and non-parametric models connected to higher order asymptotics
- Effective dimension reduction for inference
- Ensemble methods in machine learning



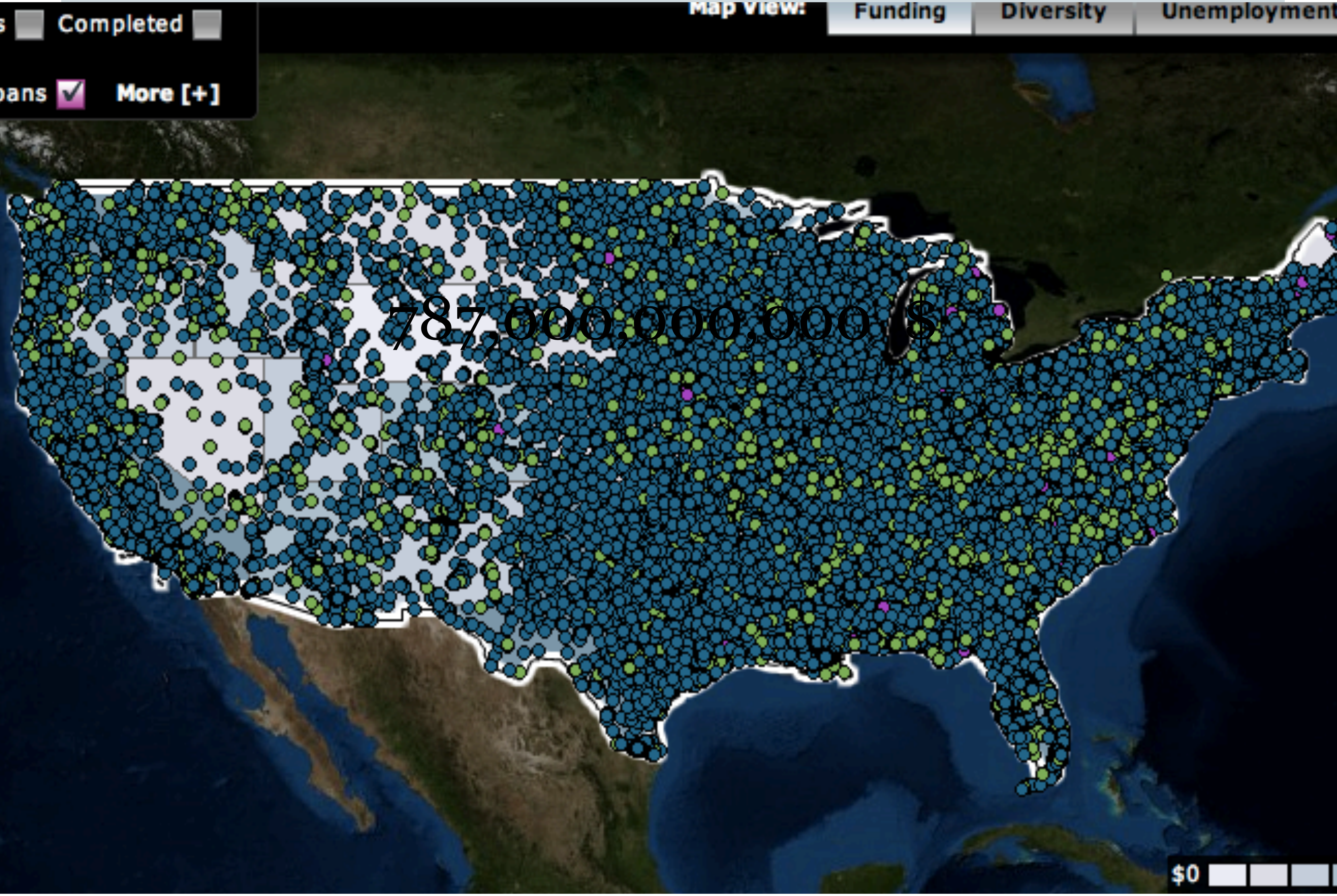
Speculation

- “in statistics the problems always evolve relative to the development of new data structures and new computational tools” … NSF report
- “Statistics is driven by data” … Don McLeish
- “Our discipline needs collaborations” … Hugh Chipman
- How do we create opportunities?
- How do we establish an independent identity?
- In the face of bureaucratic pressures to merge?
- Keep emphasizing what we do best!!



Speculation

- Engle
  - Variation, modelling, data, theory, data, theory
- Tibshirani
  - Cross-validation; forensic statistics
- Netflix Grand Prize
  - Recommender systems: machine learning, psychology, statistics!
- Tufte
  - "Visual Display of Quantitative Information" -- 1983



787,000,000,000 $

Thank you!!

2010 Annual Meeting in Québec City

38th Annual Meeting of the Statistical Society of Canada

End Notes

1. "Making Sense of Statistics" Accessed on May 5, 2010. http://www.senseaboutscience.org.uk/
2. Midlife Crisis: National Post, January 30, 2008.
3. Alessandra Brazzale, Anthony Davison and Reid (2007). *Applied Asymptotics*. Cambridge University Press.
4. Amari (1982). *Biometrika*.
5. Fraser, Reid, Jianrong Wu. (1999). *Biometrika*.
6. Reid (2003). *Annals Statistics*
7. Fraser (1990). *J. Multivariate Anal.*
8. Figure drawn by Alessandra Brazzale. From Reid (2003).
9. Davison, Fraser, Reid (2006). *JRSS B*.
10. Davison, Fraser, Reid, Nicola Sartori (2010). in progress
11. Reid and Fraser (2010). *Biometrika*
12. Fraser, Reid, Elisabetta Marras, Grace Yun-Yi (2010). *JRSSB*
13. Reid and Ye Sun (2009). *Communications in Statistics*
14. J. Heinrich (2003). *Phystat Proceedings*
15. C. Varin, C. Czado (2010). *Biostatistics*.
16. D.Cox, Reid (2004). *Biometrika*.
17. CL references in C.Varin, D.Firth, Reid (2010). Submitted for publication.
18. Account of FDR and astronomy taken from Lindsay et al (2004). NSF Report on the Future of Statistics
19. Miller et al. (2001). *Science*.
20. Photo: http://epiac1216.wordpress.com/2008/09/23/origins-of-the-phrase-pie-in-the-sky/
21. Photo: http://www.bankofcanada.ca/en/banknotes/legislation/images/023361-lg.jpg