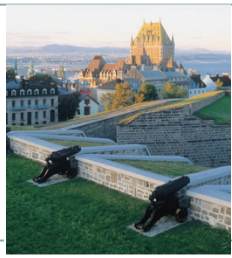
Pensées sur la théorie statistique

Nancy Reid

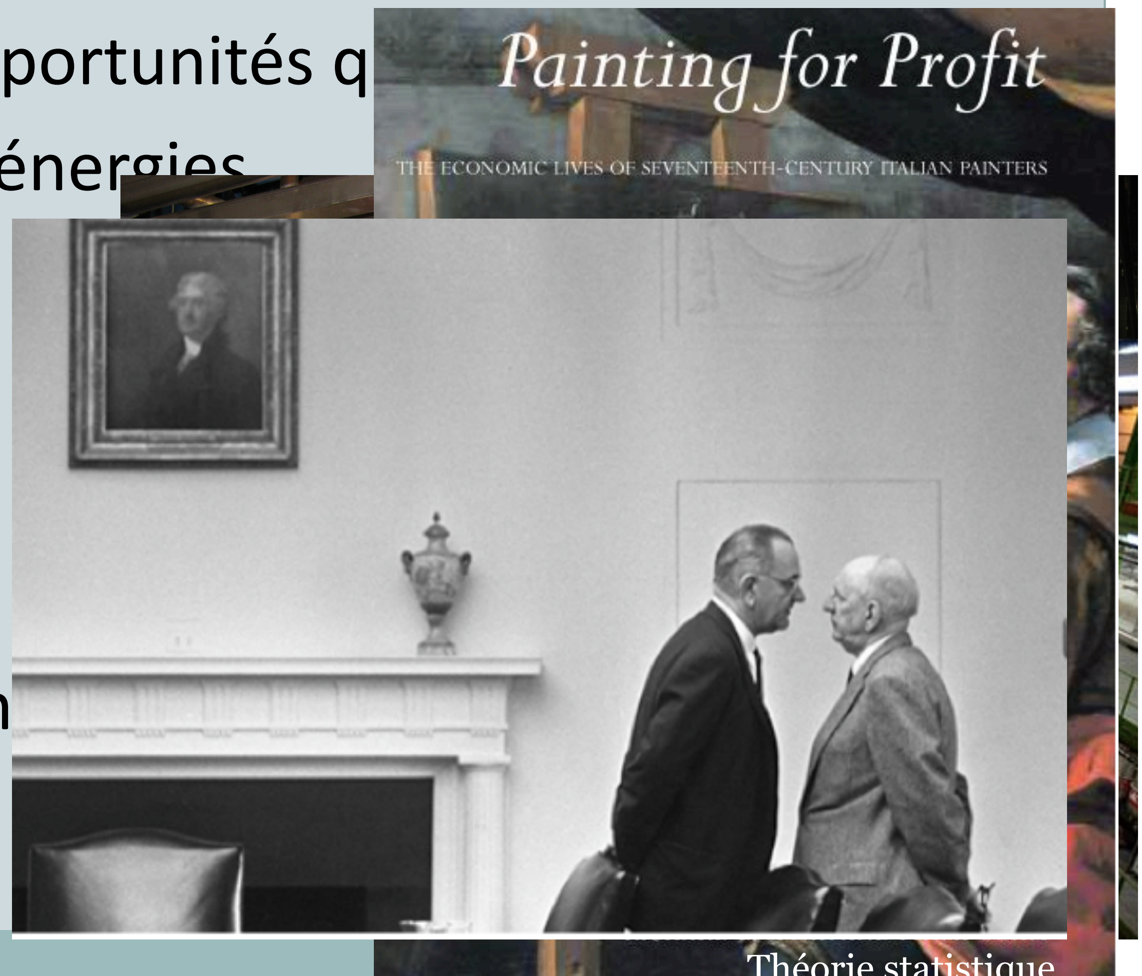
Congrès Annuel 2010 à Québec

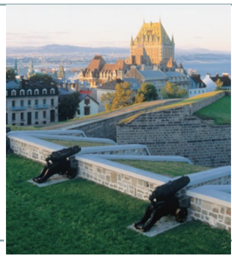
38e Congrès annuel de la Société statistique du Canada



La statistique en demande

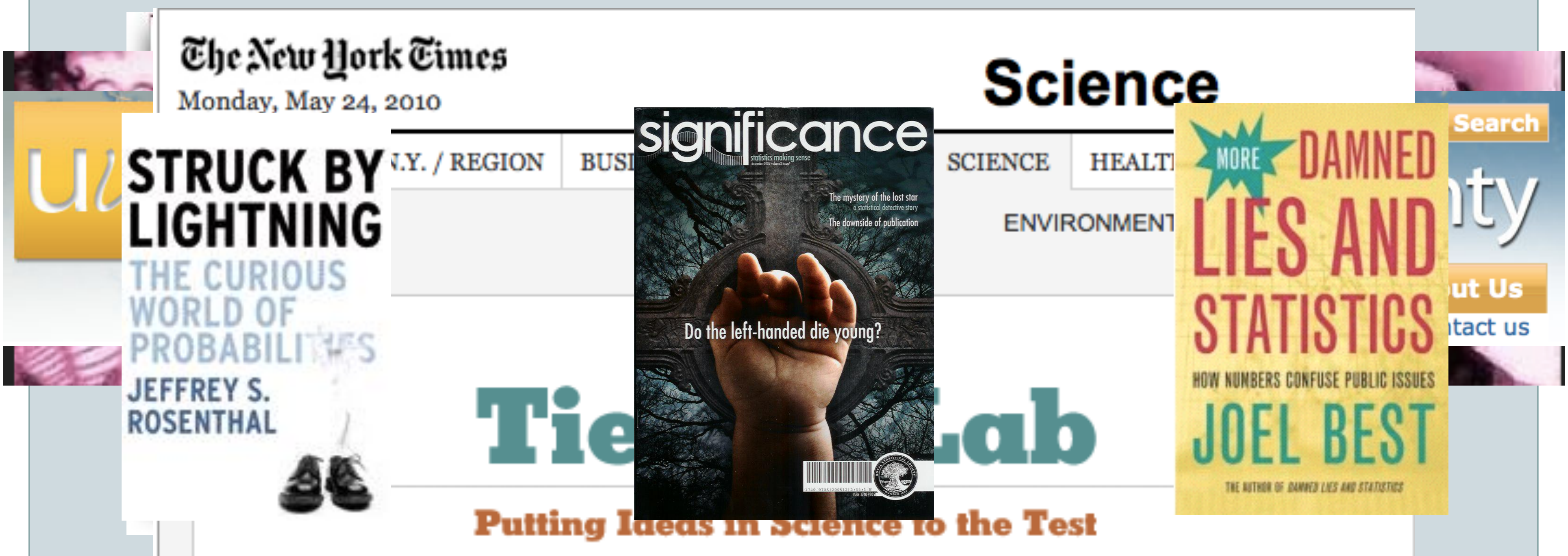
- “La science statistique connaît une croissance sans précédent tant en opportunités q
- Physique des hautes énergies
- Histoire de l'art
- Forage de réalité
- Bioinformatique
- Enquêtes complexes
- Climat et environnem
- SSC 2010 …

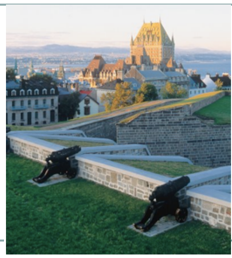




Pensée Statistique

- Pleins des ressources



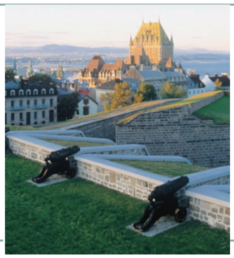


Pensée statistique

- Si une statistique était la réponse, quelle est la

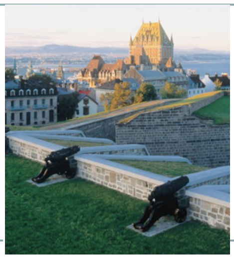
SENSE ABOUT SCIENCE
AND STRAIGHT STATISTICS
MAKING SENSE OF STATISTICS

  - signification et confiance statistiques
- Pourcentages et risques
  - changements relatifs et absolus



Théorie statistique pour 20xx

- Que devrions-nous enseigner?
- Si une statistique est la réponse, quelle est la question?
  - Planification d'expérience et enquêtes
- Embûches fréquentes
  - Statistiques descriptives: exhaustivité etc.
- Sommes-nous certains?
  - Inférence
- Pourcentages et risques
  - Interprétation



Modèles et vraisemblance

- La modélisation est difficile et importante
- Il y a beaucoup à tirer de la fonction de vraisemblance
- Pas que des estimateurs ponctuels de $\hat{\theta}$
- Pas que (pas du tout!) des tests les plus puissants $\dfrac{f(y;\theta_1)}{f(y;\theta_0)}$
- Quantités inférentielles (pivots)
- Distributions inférentielles (asymptotiques)
- Un point de départ naturel, même pour des modèles très complexes

Estimating Genotypic Correlations and Their Standard Errors Using Multivariate Restricted Maximum Likelihood Estimation with SAS Proc MIXE...

file:///Users/nancy/talks/waterloo/cropscience.html

(12) **United Sta**

Coene et al.

7,058,142 B2

Jun. 6, 2006

(54) **GENERATION OF**
FOR A PARTIAL R
LIKELIHOOD (PR

(75) Inventors: **Willem M**
Renatus .
Eindhove

(73) Assignee: **Koninklij**
Eindhove

(*) Notice: Subject to
patent is
U.S.C. 15

(21) Appl. No.: **10/403,54**

d

MENTS

al.	714/795
....	371/43
et al.	714/752
....	360/32
t al.	714/703
al.	714/755
....	341/59
et al.	399/116

ael E. Belk

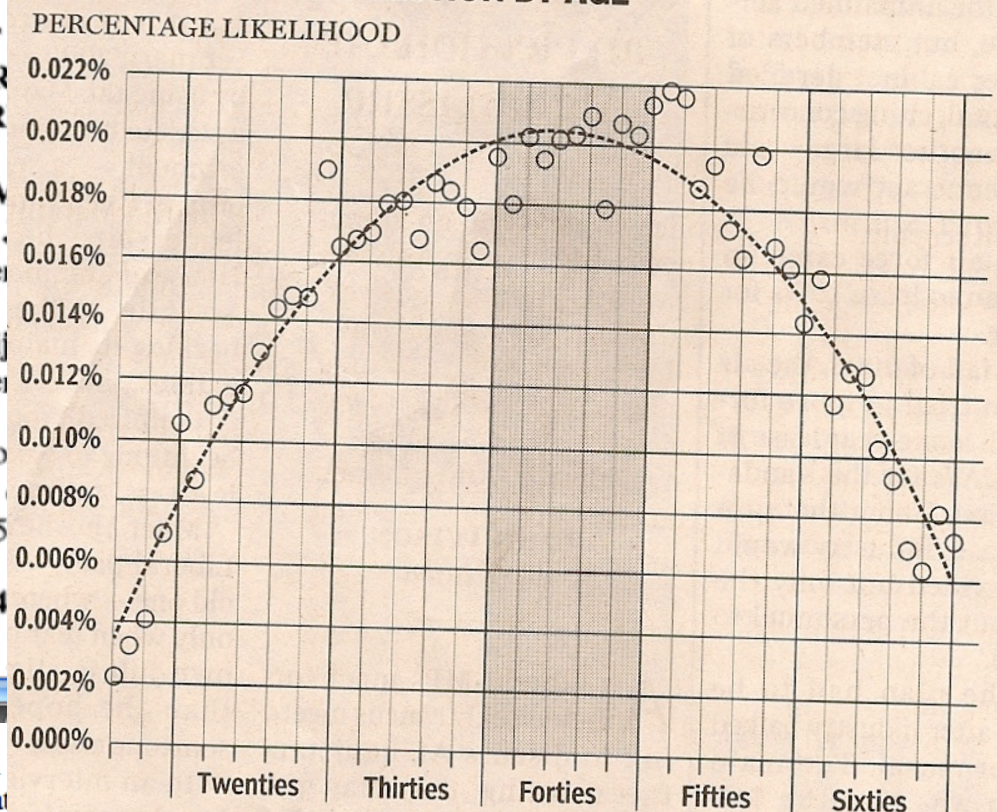
values from an input
values can he used as
state machine. which

higher power of detection of correlat
genotypes and two environments are

HAVING A MID-LIFE CRISIS? YOU'RE NOT ALONE

A study involving two million people in 72 countries found men and women were less happy in their 40s but that improved in later life.

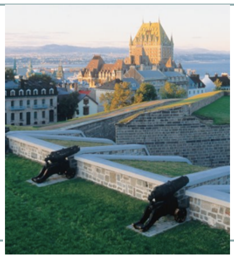
PROBABILITY OF DEPRESSION BY AGE



SOURCES: IS WELL-BEING U-SHAPED OVER THE LIFE CYCLE?

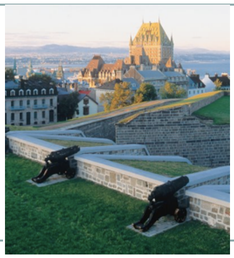
RICHARD JOHNSON / NATIONAL POST

Done

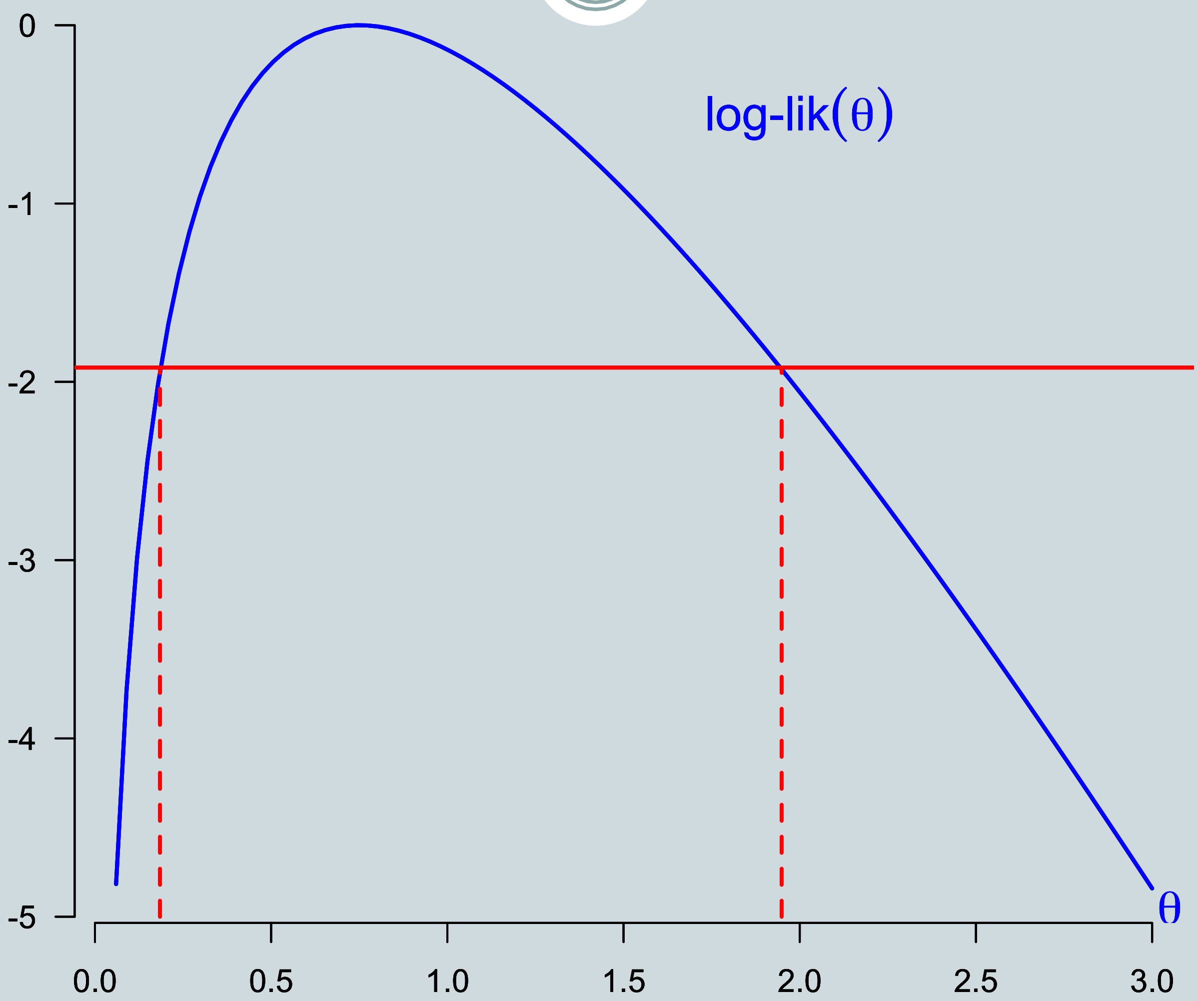


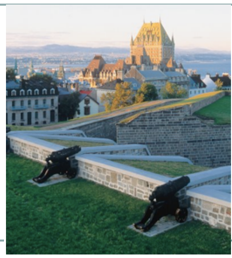
Aperçu

1. Théorie asymptotique d'ordre supérieur/la vraisemblance en tant que pivot
2. Inférence bayésienne et non bayésienne
3. Vraisemblance partielle, quasi et composite
4. Où allons-nous?

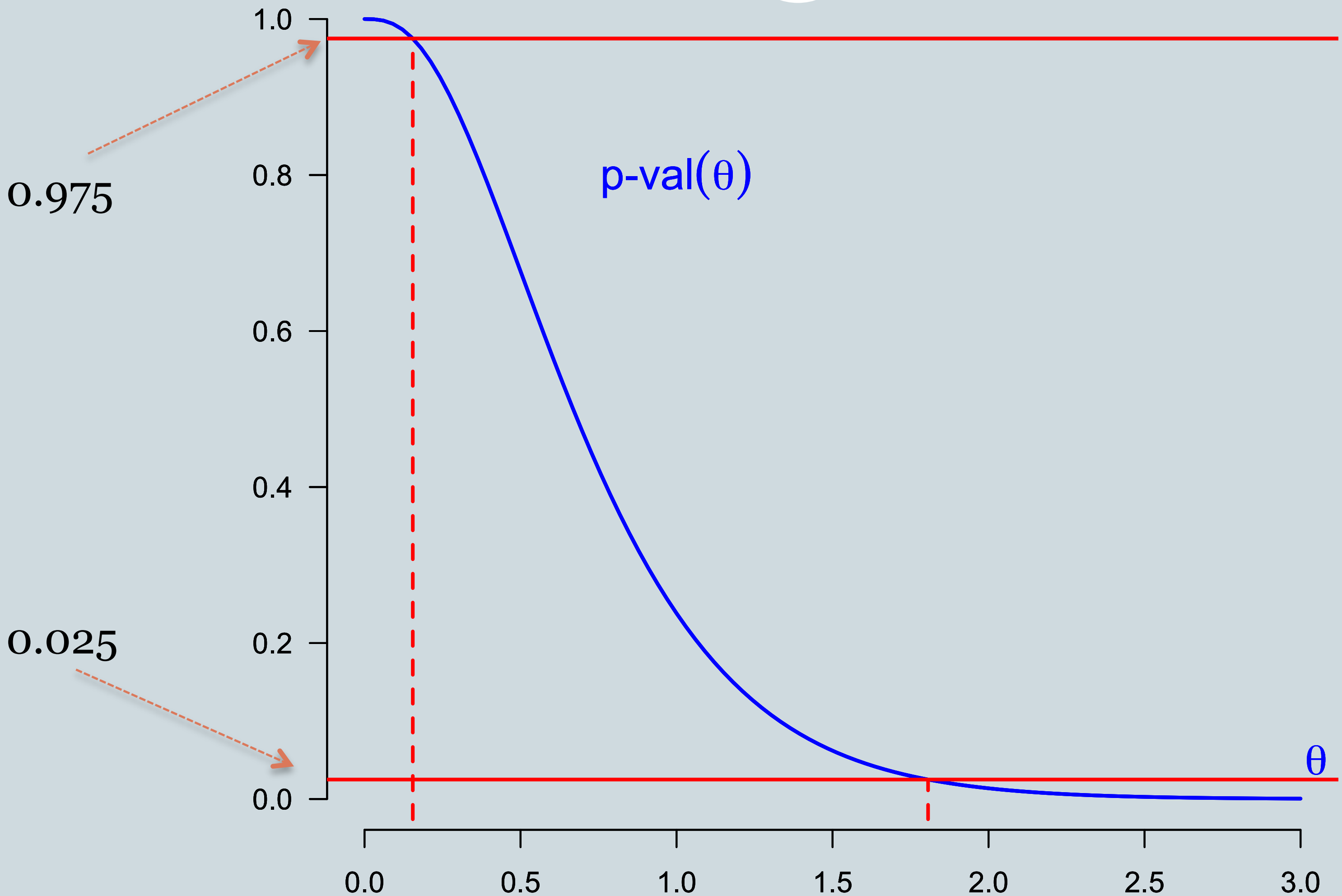


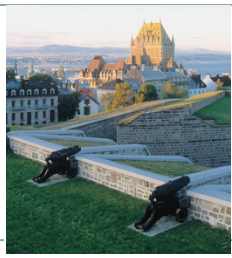
Fonctions valeur-p à partir de vraisemblance





Fonctions valeur-p à partir de vraisemblance



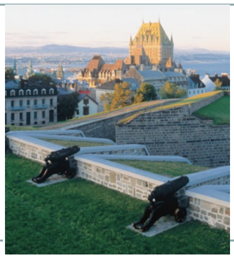


Peut être presque exacte

- Racine de la vraisemblance $r(\theta) = \pm\sqrt{[2\{\ell(\hat{\theta}) - \ell(\theta)\}]}$
- Estimateur du max. de vraisemblance $q(\theta) = (\hat{\theta} - \theta) j^{1/2}(\hat{\theta})$
- Fonction de score $s(\theta) = \ell'(\theta) j^{-1/2}(\hat{\theta})$
- Toutes approximativement de loi $N(0,1)$

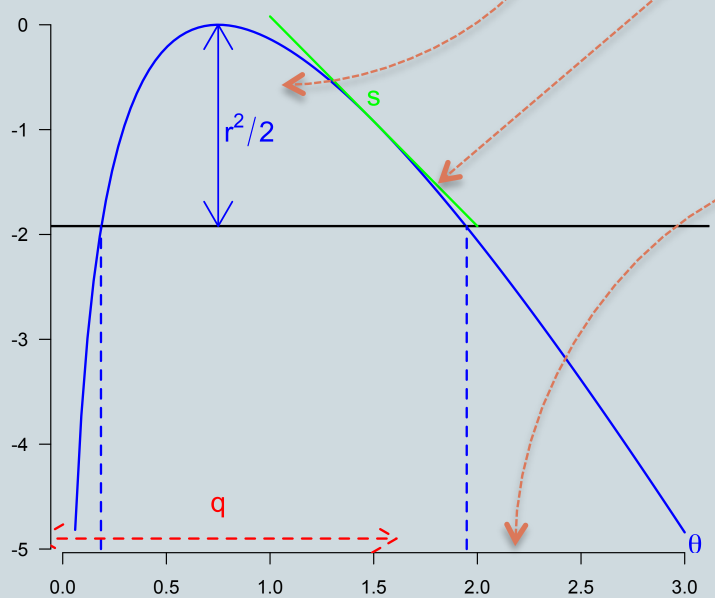
Beaucoup mieux: $r^*(\theta) = r(\theta) + \frac{1}{r(\theta)} \log \frac{Q(\theta)}{r(\theta)}$

- $Q(\theta)$ peut être $q(\theta)$ or $s(\theta)$ or ...



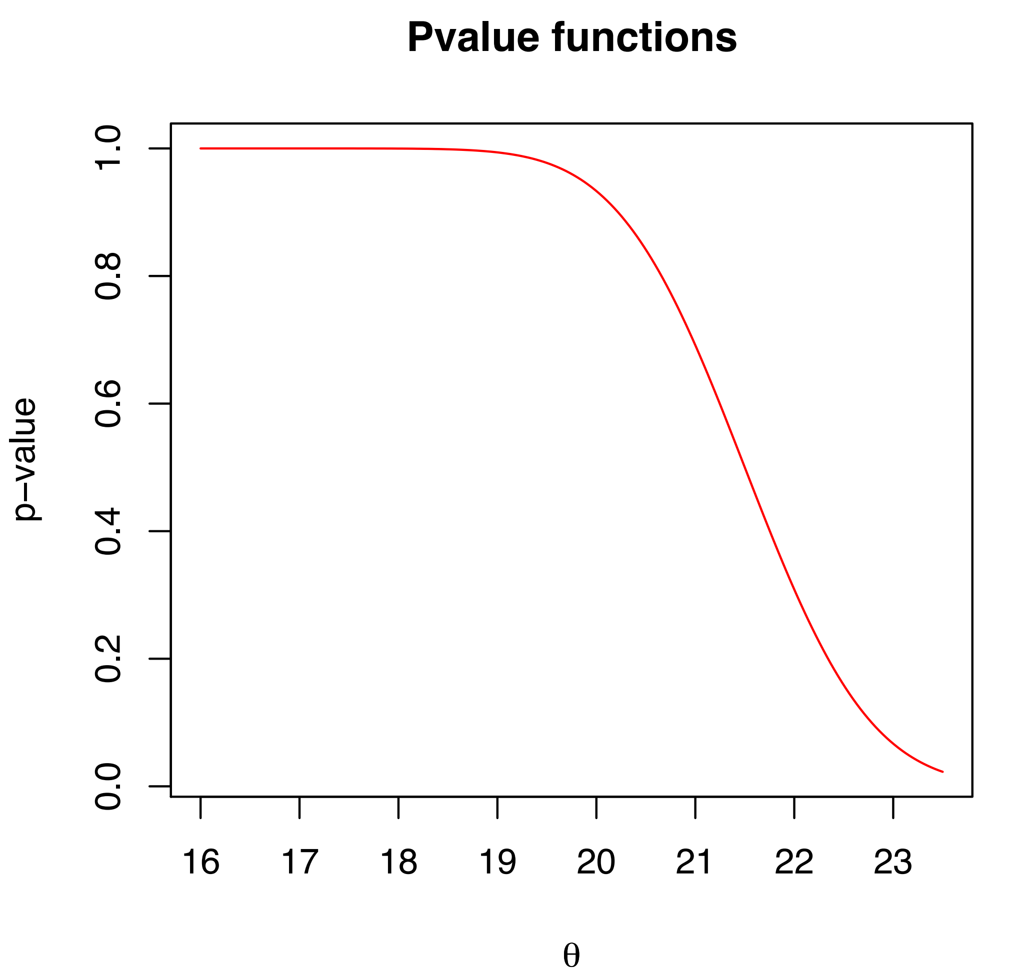
Peut être presque exacte

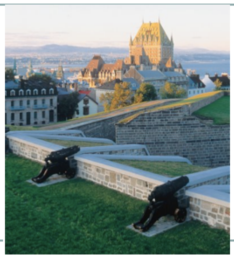
- Racine de la vraisemblance $r(\theta) = \pm\sqrt{[2\{\ell(\hat\theta) - \ell(\theta)\}]}$
- Estimateur du max. de vraisemblance $q(\theta) = (\hat\theta - \theta) j^{1/2}(\hat\theta)$
- Fonction de score $s(\theta) = \ell'(\theta) j^{-1/2}(\hat\theta)$



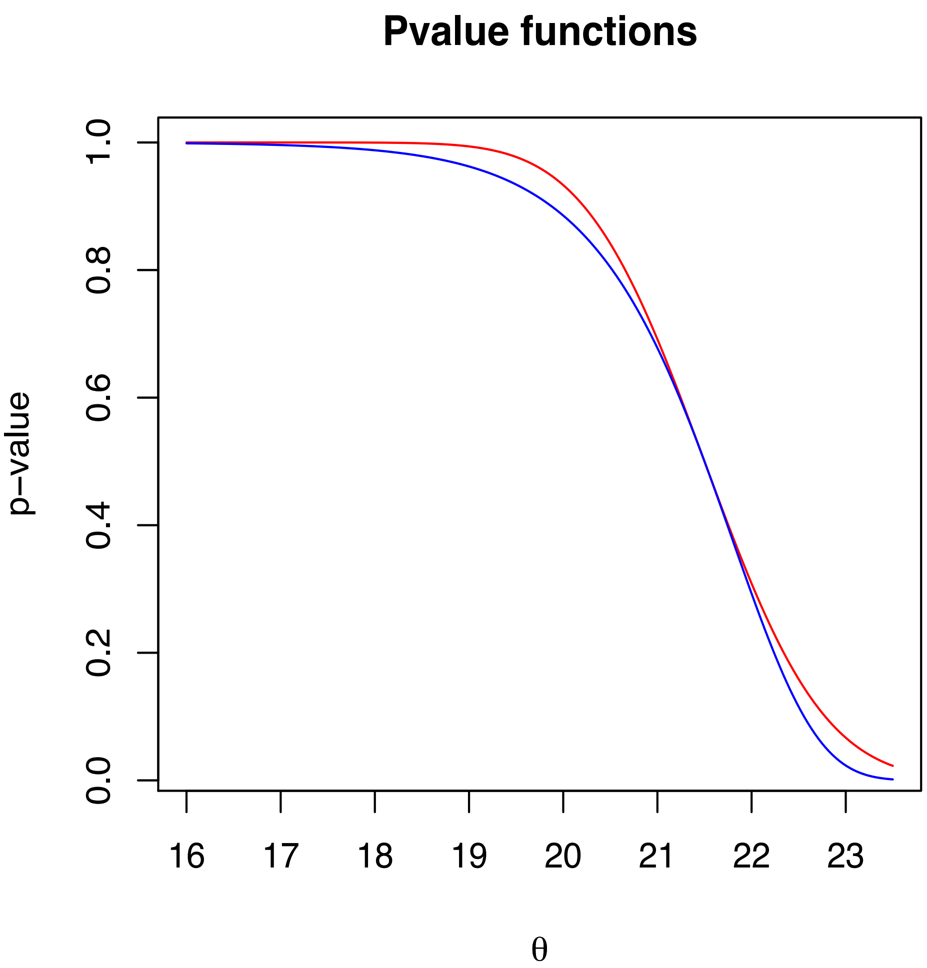
$$j(\theta) = -\ell''(\theta)$$

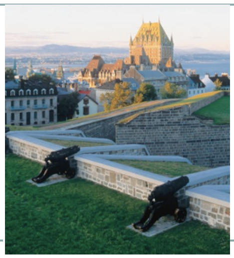
Peut être presque exacte



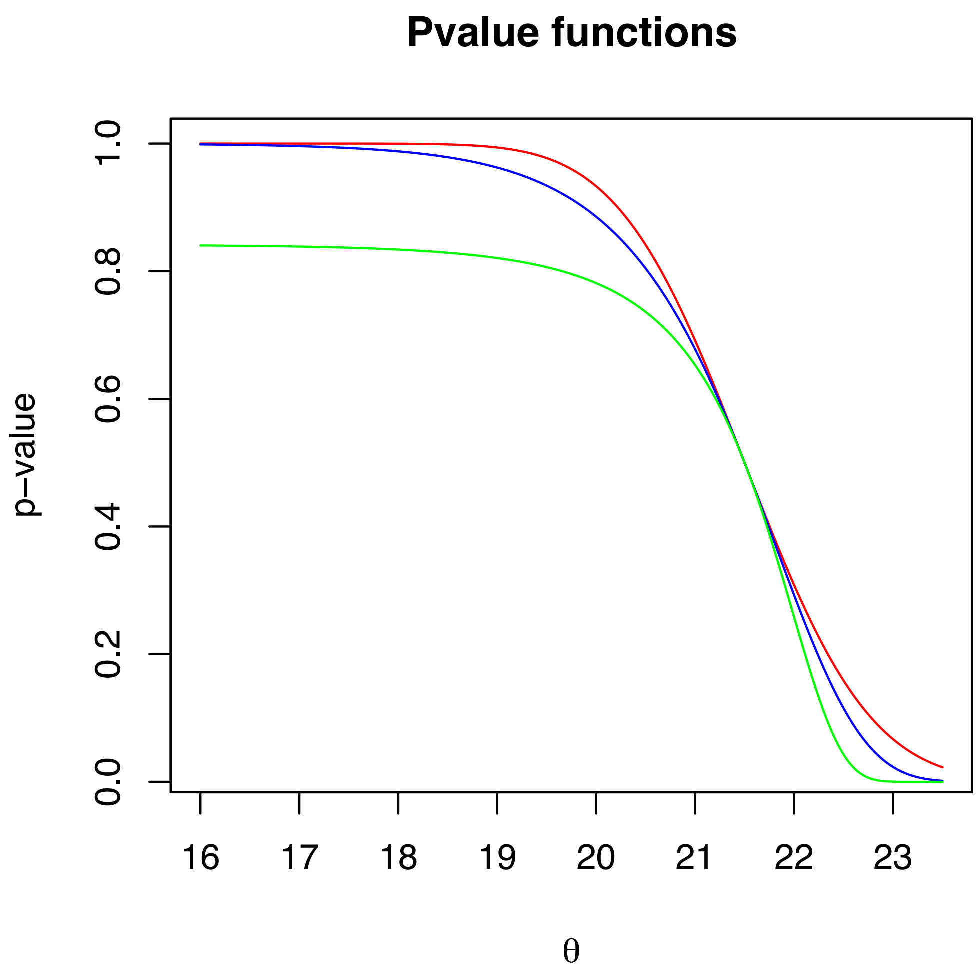


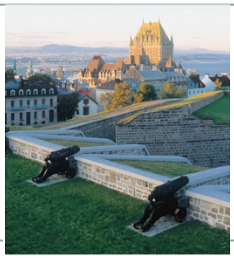
Peut être presque exacte



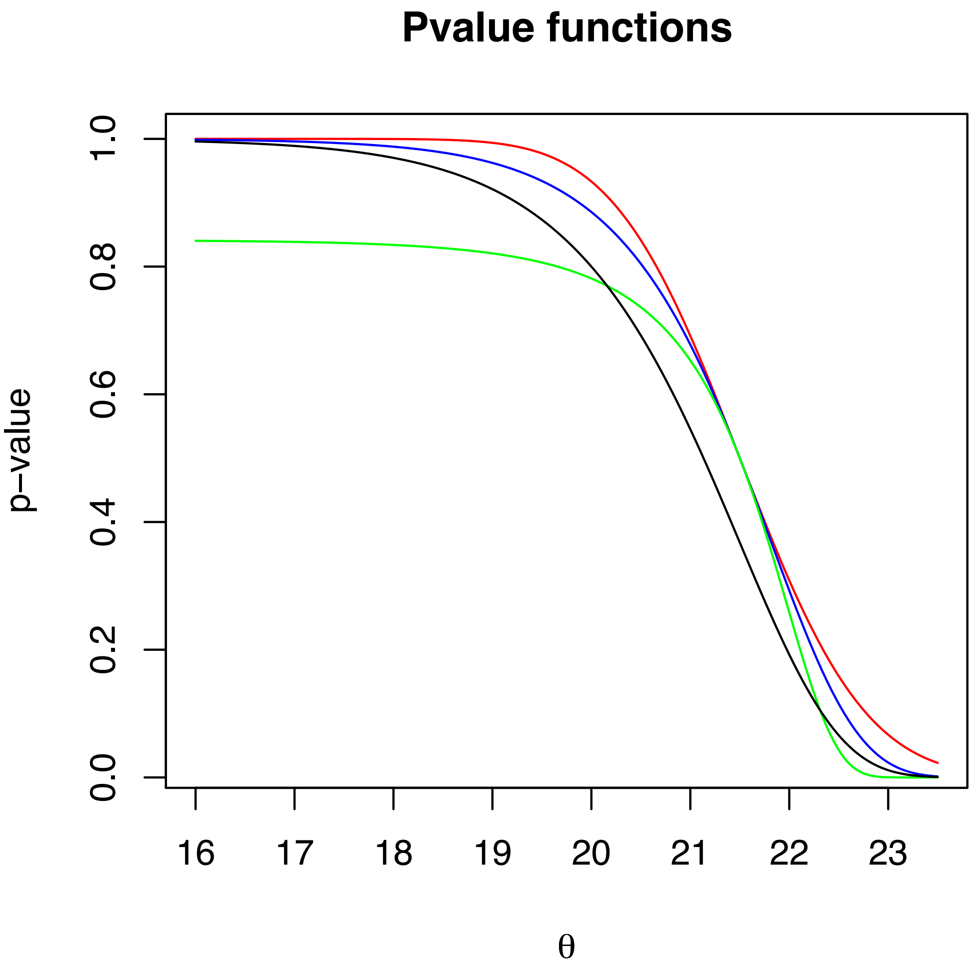


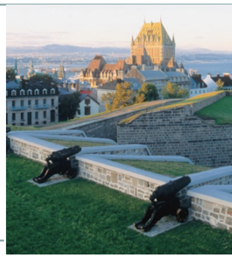
Peut être presque exacte





Peut être presque exacte

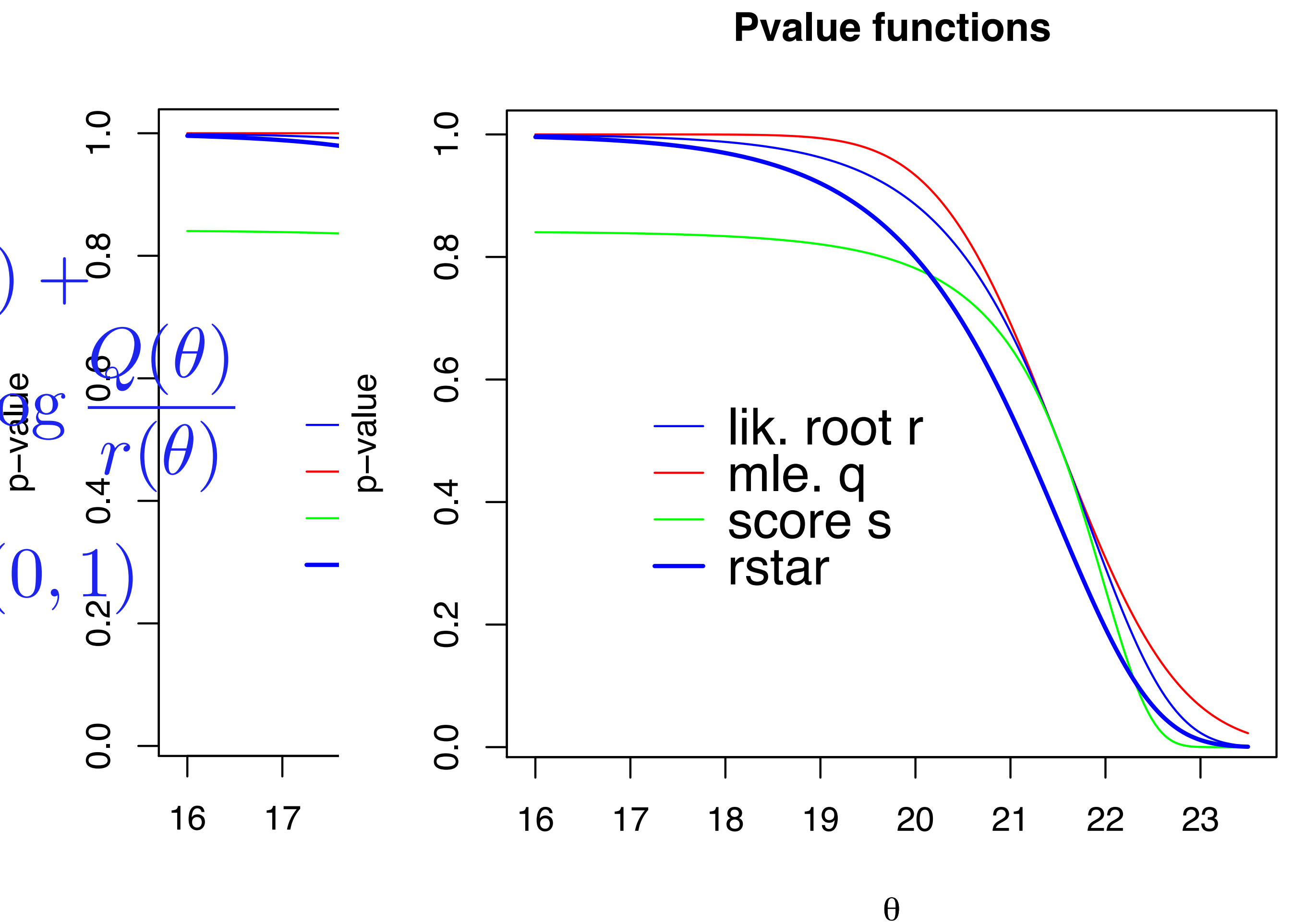


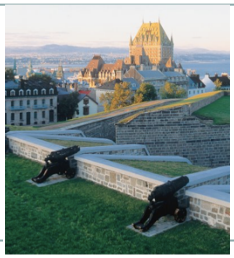


Peut être presque exacte

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

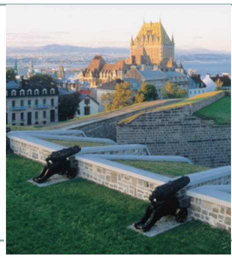
$$\simeq N(0,1)$$





Utiliser des approximations d'ordre supérieur

- Approximations excellentes pour cas 'faciles'
  - Familles exponentielles, régression linéaire non normale
- Demandant plus de travail pour cas 'modérés'
  - Modèles autorégressifs, effets fixes et aléatoires, réponse discrète
- Délicates pour cas 'difficiles'
  - Modèles structurels complexes à sources de variation multiples
- Meilleurs résultats pour un paramètre scalaire
  - Mais on peut devoir faire de l'inférence sur des paramètres vectoriels



D’où est-ce que ça vient?

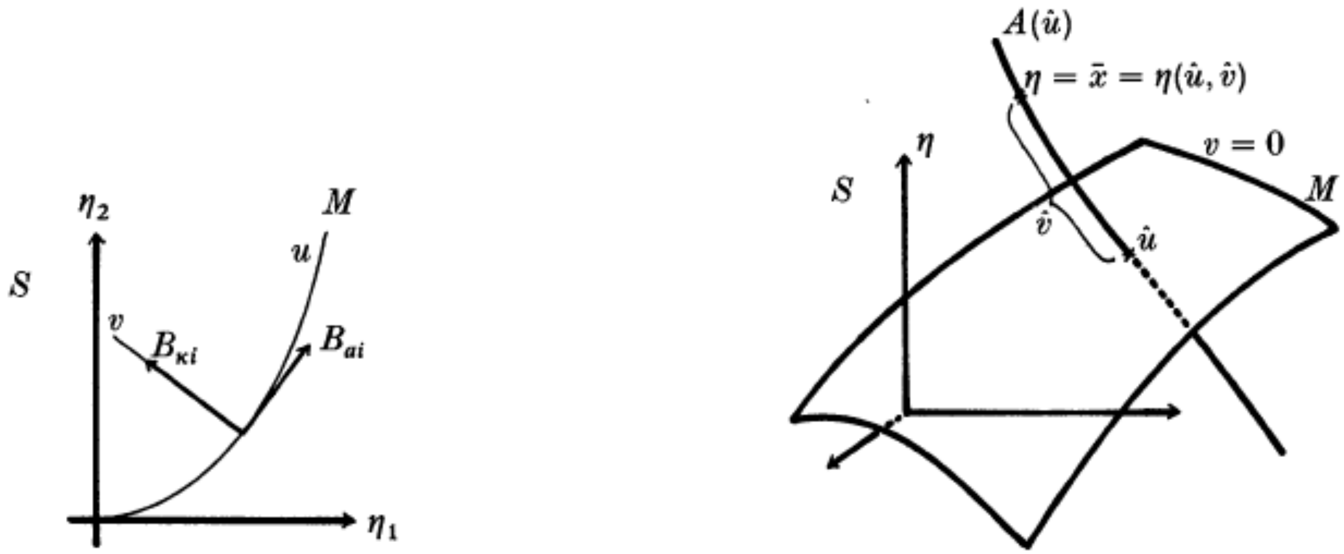
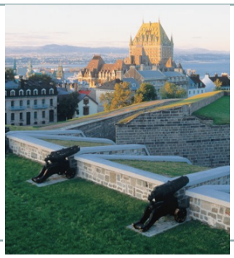


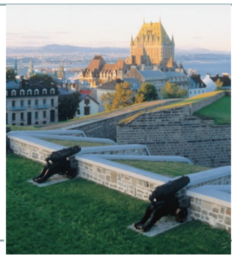
Fig. 1 (left). Example of curved exponential family $N(u, a^2 u^2)$.
Fig. 2 (right). Ancillary subspace $A(u)$ and local coordinates (u, v).

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



D'où est-ce que ça vient?

- Géométrie différentielle de modèles statistiques
- Théorie des familles exponentielles
- Approximations d'Edgeworth et du point de selle
- Idée clef:
- Un modèle paramétrique lisse peut être approximé par un modèle tangent tiré d'une famille exponentielle
- Nécessite de différencier la fonction de log-vraisemblance par rapport à l'espace échantillonnal
- Permet des généralisations aux modèles plus complexes



D'où est-ce que ça vient?

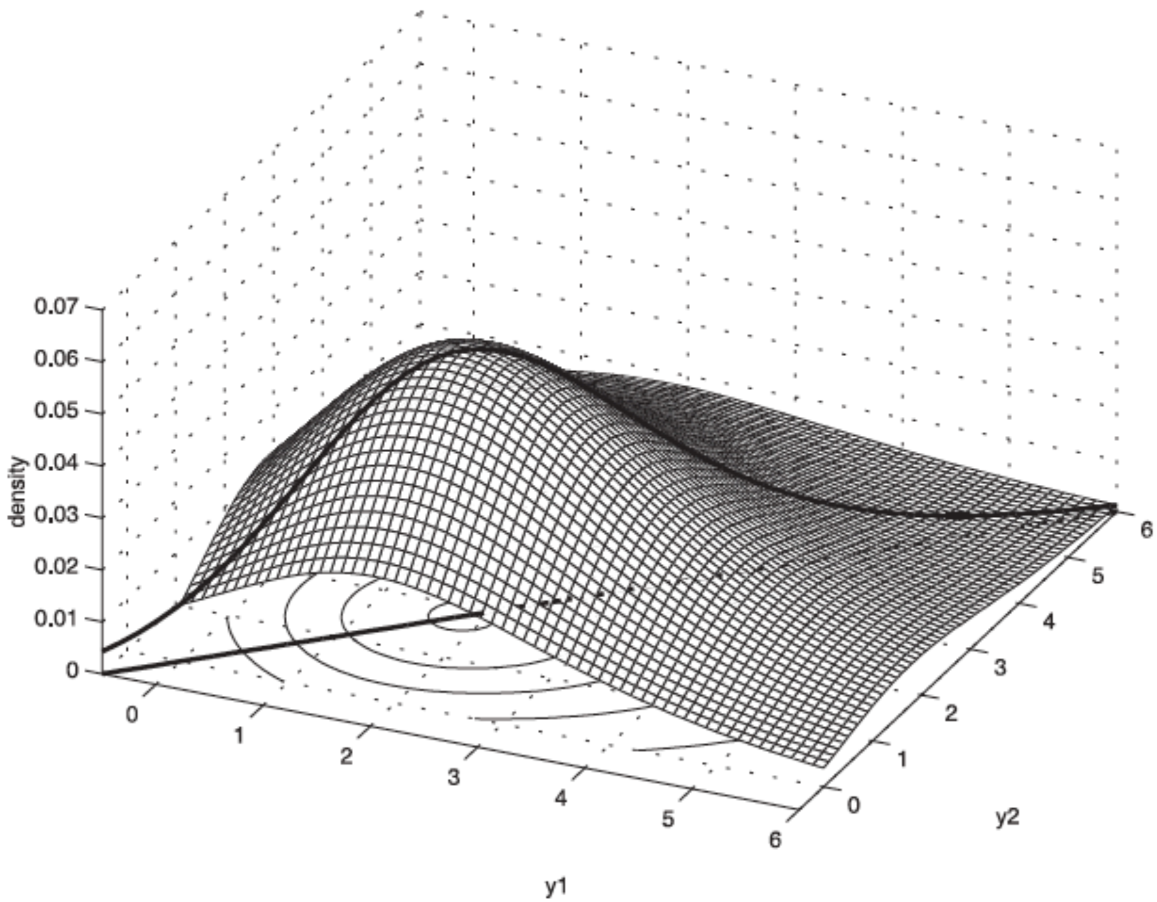
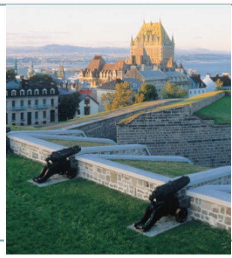


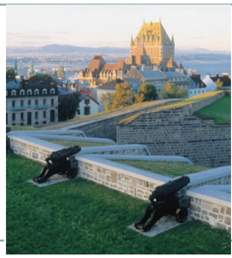
FIG. 2. *The second-order ancillary, with tangent vectors given by V, is constant along the solid curve in the (y_1, y_2) plane.*

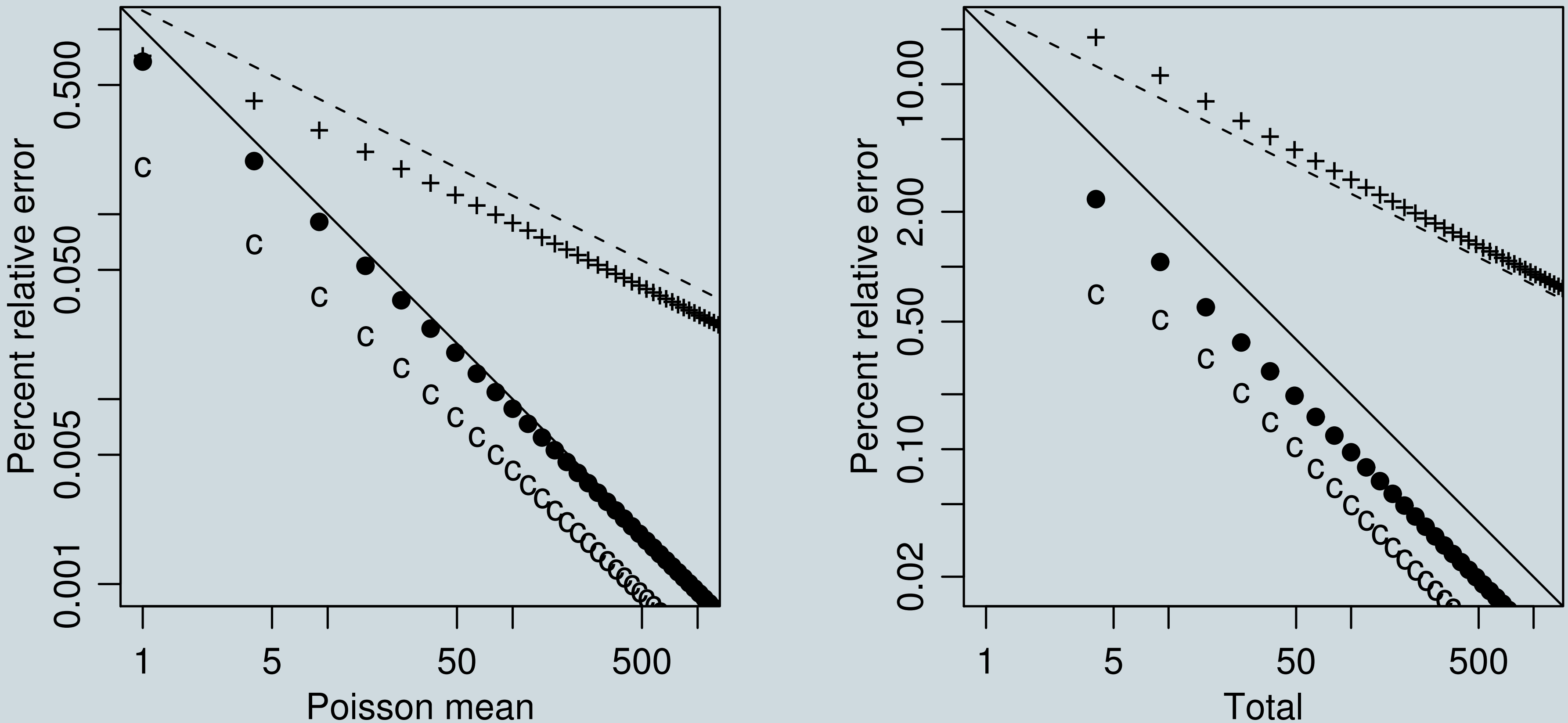


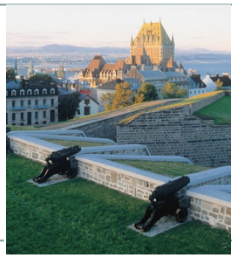
Généralisations

- À des données discrètes
- Lorsque différencier la log-vraisemblance par rapport à l'espace échantillonnal est plus
- Solution: utiliser plutôt la valeur espérée de la statistique de score
- L'erreur relative est alors $O(n^{-1})$ plutôt que $O(n^{-3/2})$
- Tout de même mieux que l'approximation normale

Généralisations

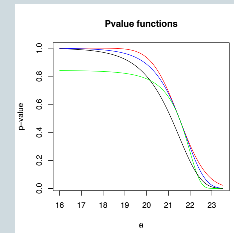


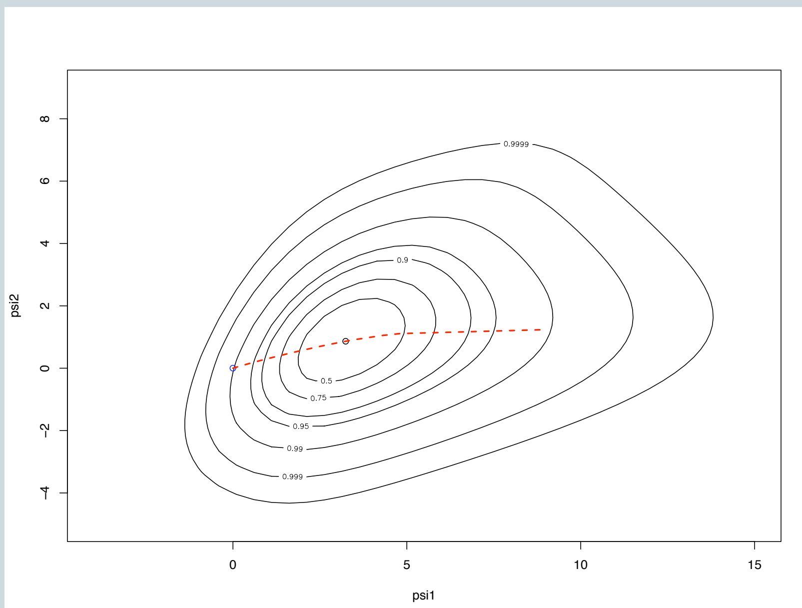


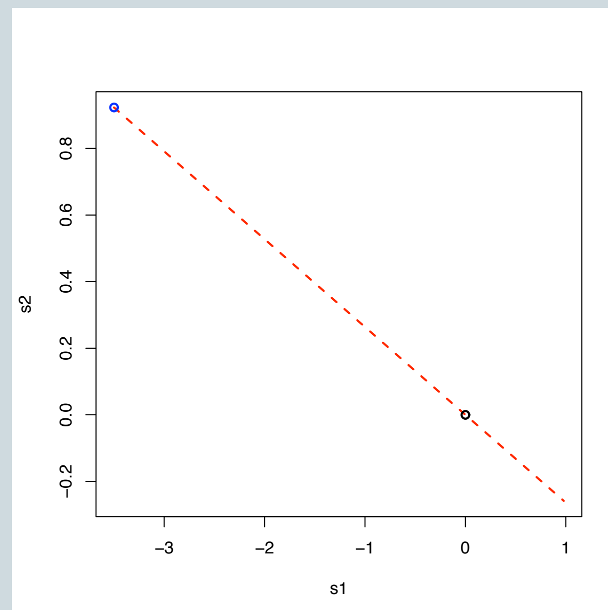


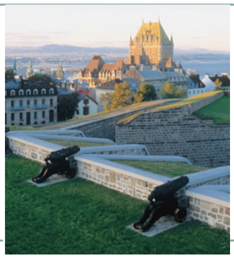
Généralisations

- À des paramètres d'intérêt vectoriels
- Mais nos solutions nécessitent un seul paramètre
- Solution: utiliser la longueur du vecteur, conditionnelle à sa direction



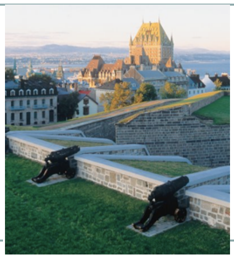






Généralisations

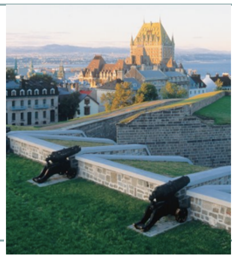
- Étendre le rôle de la famille exponentielle
- En généralisant la différentiation par rapport à l'espace échantillonnal
- Idée: différencier la log-vraisemblance espérée
  - Plutôt que la log-vraisemblance
- Mène à une nouvelle version de la famille exponentielle approximante
- Peut être utilisée avec les pseudo-vraisemblances



Que pouvons-nous apprendre?

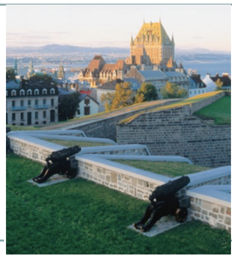
- Les approximations d'ordre supérieur demandent
- De différencier la fonction de log-vraisemblance

  par rapport à l'espace échantillonnal
- L'inférence bayésienne sera différente
- Les développements asymptotiques soulignent cet écart
- Les *a posteriori* bayésiens ne sont en général pas calibrés
- Ne peuvent pas toujours être corrigés par le choix d'*a priori*
- Nous pouvons étudier ceci en comparant approximations bayésiennes et non bayésiennes



Exemple: inférence pour ED50

- Régression logistique avec une seule covariable
- Sur l'échelle logistique $\Pr(y_i = 1) = \alpha + \beta x_i$
- Lois a priori plates pour (α, β)
- Le paramètre d'intérêt est $\psi = -\alpha/\beta$
- Couverture empirique des intervalles bayésiens *a posteriori*:
  - 0.90, 0.88, 0.89, 0.90
- Couverture empirique d'intervalles utilisant $\Phi(r^*)$
  - 0.95, 0.95, 0.95, 0.95



Loi *a priori* plates: mauvaise idée!

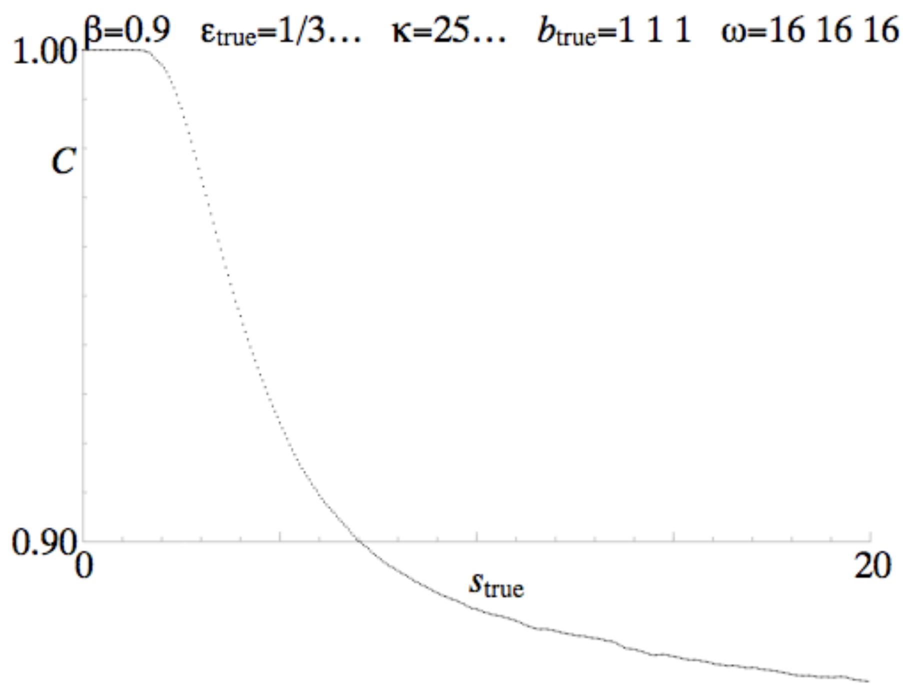
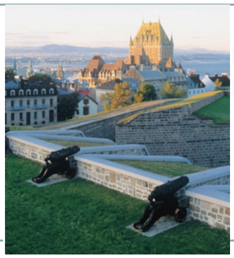
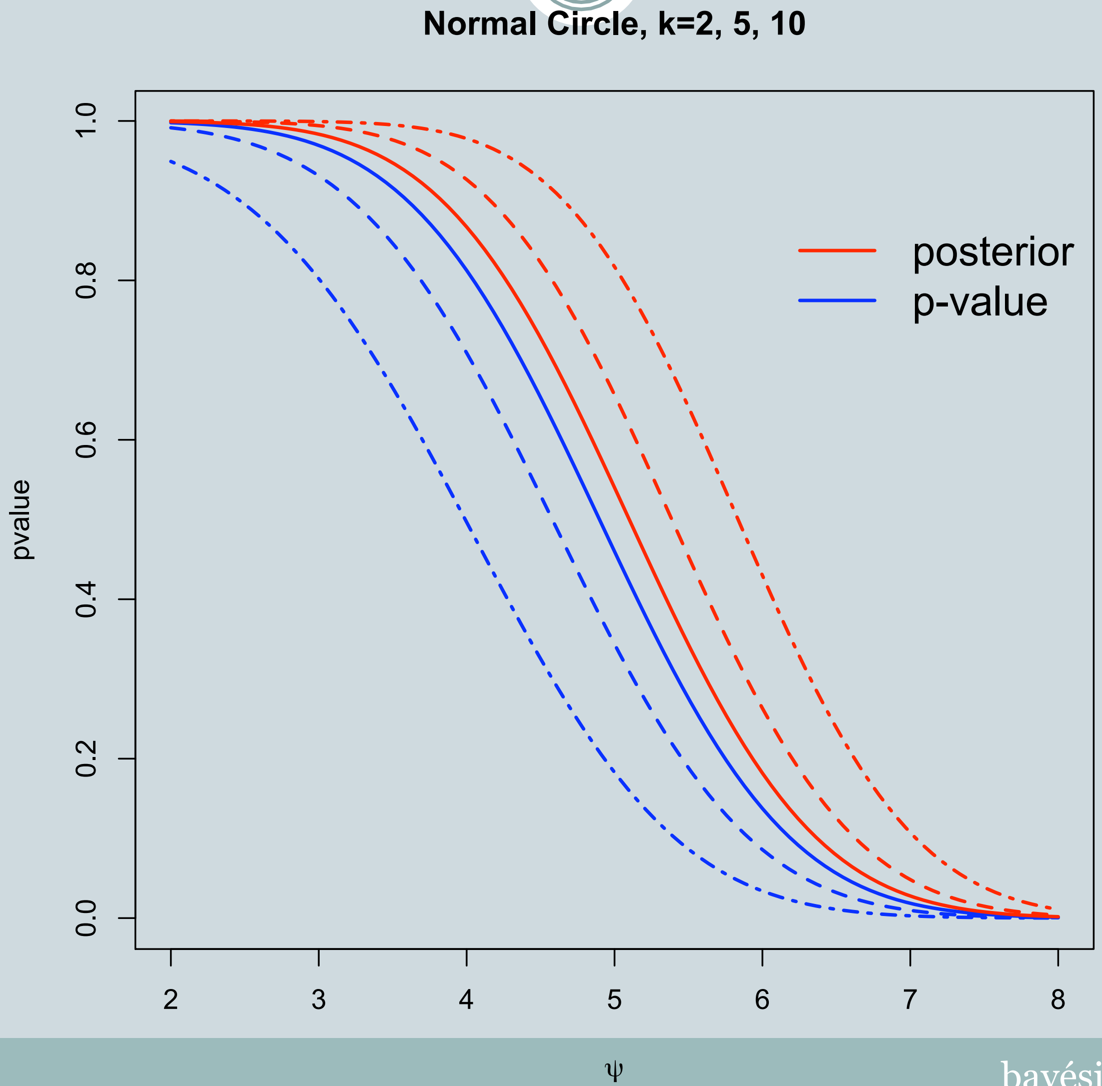
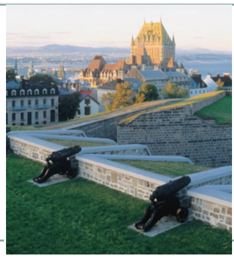


Fig. 6. 3 independent channels. Coverage for 90% credibility level upper limits, acceptance uncertainty = 34%/channel, background uncertainty = 25%/channel.



Loi *a priori* plates: mauvaise idée!





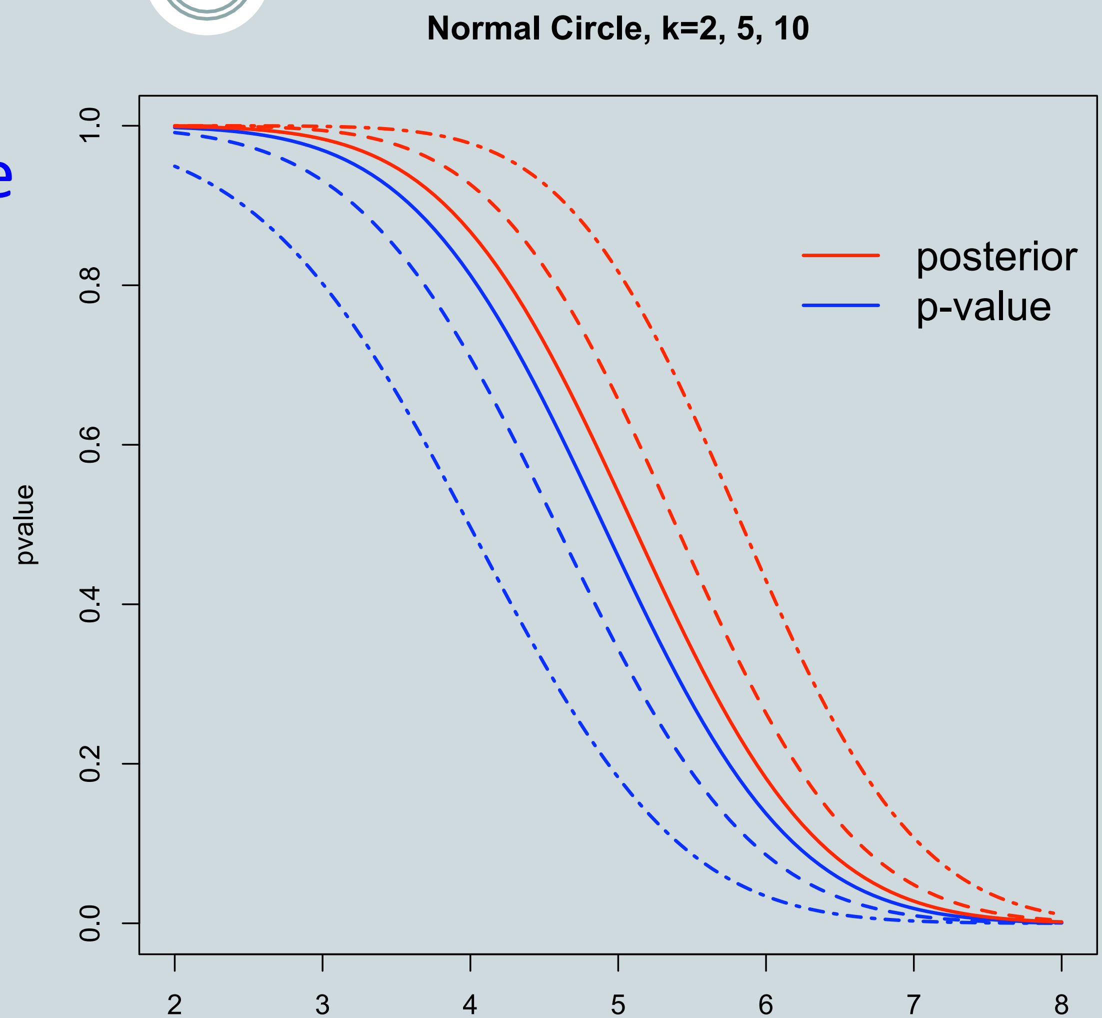
Loi *a priori* plates: mauvaise idée!

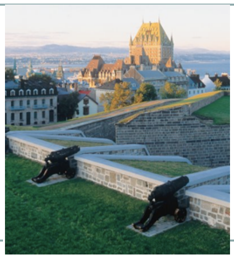
Valeur-p bayésienne – valeur-p fréquentiste

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

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

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

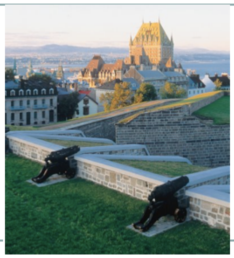




Modèles plus complexes

- L'inférence basée sur la vraisemblance a plusieurs qualités souhaitables
- Exhaustivité, efficacité asymptotique
- Bonnes approximations pour les distributions requises
- Obtenue naturellement à partir de modèles paramétriques
- Peut être difficile à construire,

  surtout pour des modèles complexes
- Plusieurs généralisations naturelles: vraisemblance partielle pour données censurées, quasi-vraisemblance pour équations d'estimation généralisées, vraisemblance composite pour données dépendantes

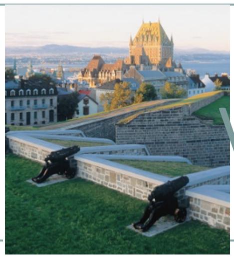


Modèles complexes

- Exemple: études longitudinales de sujets atteints de migraines
- Variable latente $Y_{ij}^* = x_{ij}^T\beta + U_i + \epsilon_{ij}$
- Variable observée

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

- E.g. maux de tête intenses, modérés, faibles, absents…
- x_{ij} Covariables: âge, éducation, analgésiques, météo, …
- U_i, ϵ_{ij} effets aléatoires inter- et intra-sujets
- Corrélation sérielle $\epsilon_{ij} = \rho\epsilon_{i,j-1} + (1-\rho^2)^{1/2}\eta_{ij}$



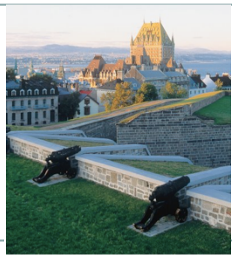
Vraisemblance pour données discrètes longitudinales

- Fonction de vraisemblance

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

- Difficile à calculer
- Hypothèses fortes
- Proposition: utiliser des densités bivariées marginales

  à la place des densités normales multivariées
- Ce qui donne un modèle mal spécifié

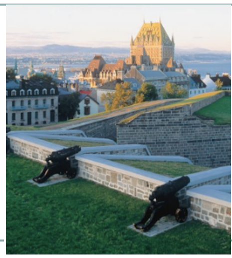


Vraisemblance composite

- Fonction de vraisemblance composite

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

- Plus généralement, $$CL(\theta) = \prod_{i=1}^{n} \prod_{k=1}^{K} f(y_i \in \mathcal{A}_k)$$
- Les ensembles $\mathcal{A}_k$ indexent des distributions marginales ou conditionnelles (ou …)
- Inférence basée sur la théorie des équations d'estimation

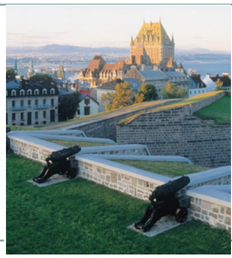


Un exemple simple

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

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

- L'EMV par paires de θ est entièrement efficace
- Si $\sigma^2 = 1$, la perte d'efficacité dépend de la dimension p
- Petite pour dimension plus petite que, disons, 10
- S'écroule si $p \to \infty$ et que la taille échantillonnale est fixe
  - Pertinent pour séries chronologiques, applications génétiques



Estimateur par vraisemblance composite

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

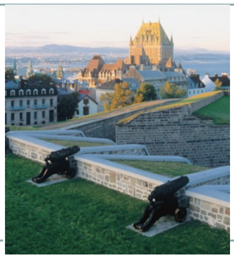
$$\hat{\theta}_{CL} \xrightarrow{p} \infty$$

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

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

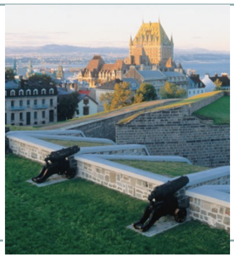
Information de Godambe

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



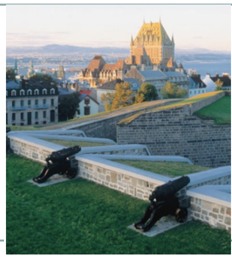
Applications récentes

- Données longitudinales binaires et continues: modèles à effets aléatoires
- Analyse de survie: modèles de défaillance, copules
- Réponses de types multiples: discrète et continue; marqueurs et temps d'évènements
- Finance: modèle à covariance variant dans le temps
- Génétique/bioinformatique: CCL pour distribution vonMises: repliement de protéine; cartographie génétique; déséquilibre de liaison
- Données spatiales: géostatistique, processus spatiaux ponctuels

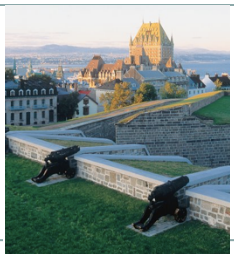


… et plus

- Analyse d’images
- Modèle de Rasch
- Modèle de Bradley-Terry
- Modèle à espace d’états
- Dynamique des populations
- …

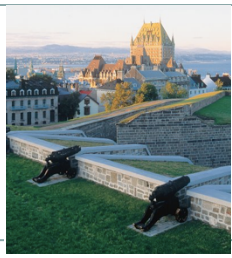


Que pouvons-nous apprendre?



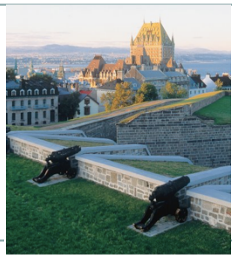
Que devons-nous savoir?

- Pourquoi les estimateurs de la vraisemblance composite sont-ils efficaces?
- Quelle quantité d'information utiliser?
- L'identifiabilité des paramètres est-elle garantie?
- Sommes-nous certains de la cohérence des composantes avec un 'vrai' modèle?
- Pouvons-nous progresser si ce n'est pas le cas?
- Comment construire les densités conjointes?
- Quelles propriétés ont ces constructions?
- La vraisemblance composite est-elle robuste?



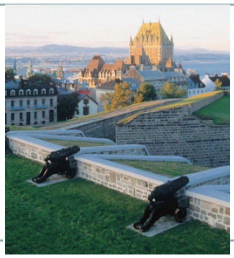
Pourquoi est-ce important?

- Les idées de la vraisemblance composite viennent des applications
- Les méthodes de vraisemblance semblent trop complexes
- Une varitété de domaines d'applications utilisent des idées similaires/identiques
- L'abstraction apportée par la théorie nous permet de prendre du recul face à l'application
- Comprendre quand les méthodes peuvent ne pas fonctionner
- Et quand devraient-elles bien fonctionner



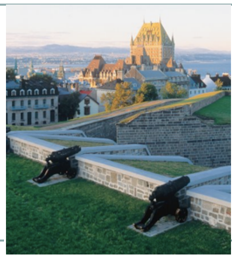
Le rôle de la théorie

- Distiller les idées principales
- Simplifier les détails
- Isoler des caractéristiques particulières
- Dans le meilleur cas, nous donner une nouvelle compréhension de ce qui est à la base de nos intuitions
- Exemple: courbure et inférence bayésienne
- Exemple: vraisemblance composite
- Exemple: taux de fausses découverte



Taux de fausses découverte

- Problème de comparaisons multiples
  - *Simultaneous statistical inference* – R.G. Miller, 1966
- La correction de Bonferroni est trop forte
- Benjamini and Hochberg, *JRSS B*, 1995
- Introduisent taux de fausses découverte
  - Beaucoup plus mieux que “Type I and Type II error”
- Et puis les données, en ce cas de l’astrophysique
- Genovese & Wasserman avec Miller & Nichol



Taux de fausses découverte

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

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

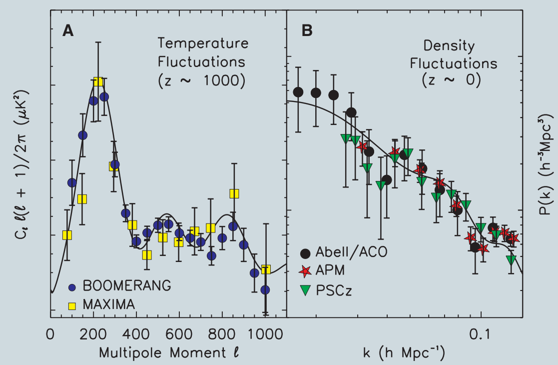
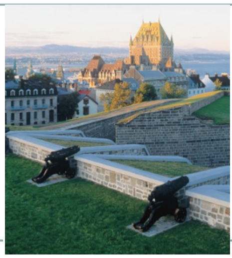
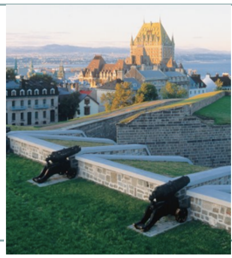


Fig. 1. We plot the CMB data from the MAXIMA and BOOMERANG experiments (**A**) alongside the matter-density data (**B**). The solid line is the best fit model (Ω_{matter} = 0.24, $\Omega_{baryons}$ = 0.06, and n_s = 1.08 with H_0 = 69) using the matter-density data alone. The amplitudes in both plots 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 Ω_{vacuum} = 0.8, consistent with the Type Ia supernovae results (*18*).



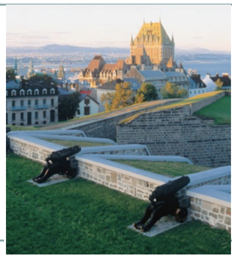
Pistes

- La vraisemblance composite comme lisseuse
- Calibration de l'inférence *a posteriori*
- Généralisation de la théorie asymptotique d'ordre supérieur à la vraisemblance composite
- Familles exponentielles et vraisemblance empirique
- Modèles semi- et non-paramétriques liés à la théorie asymptotique d'ordre supérieur
- Réduction de dimension efficace à des fins d'inférence
- Méthodes d'ensemble en apprentissage machine



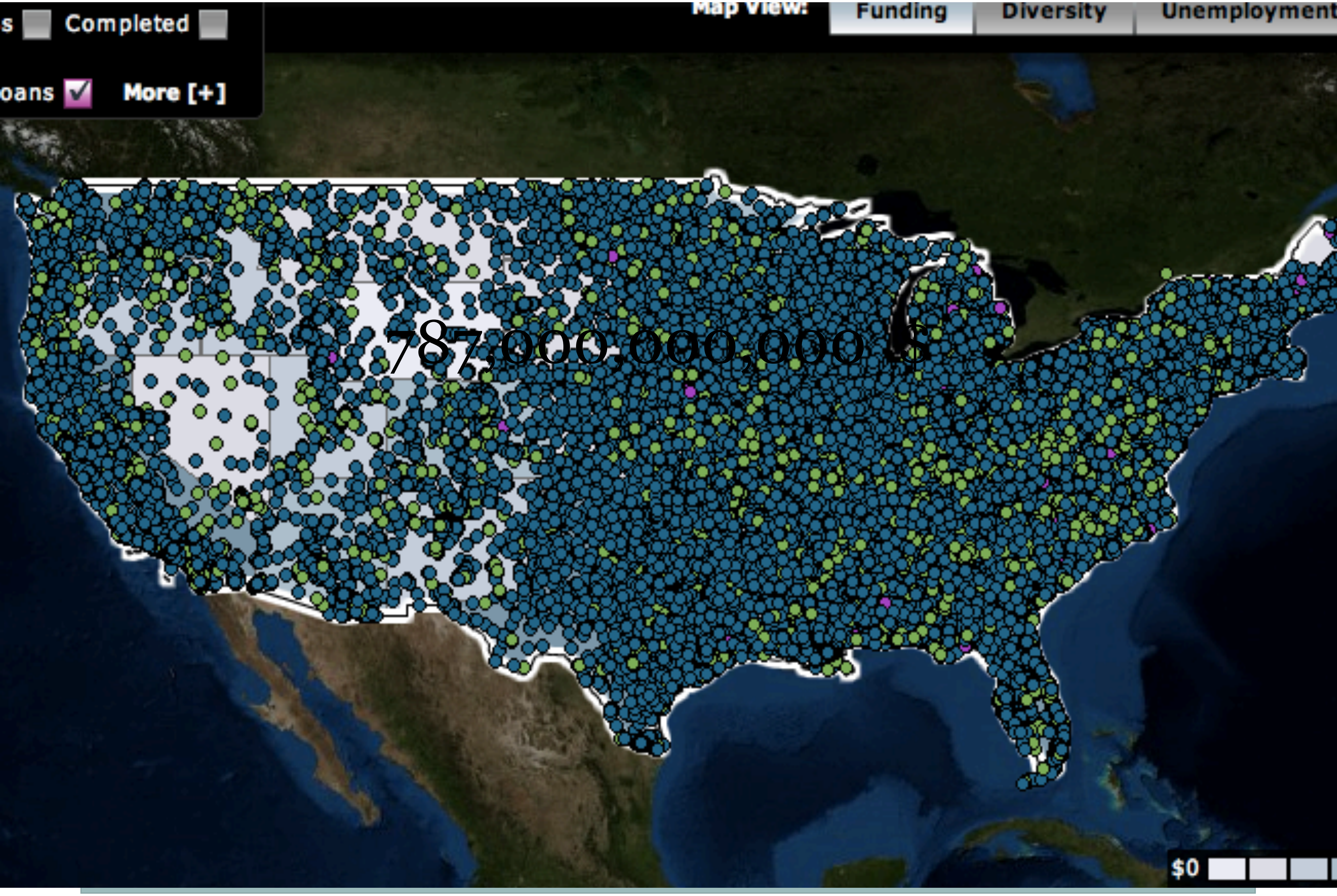
Spéculations

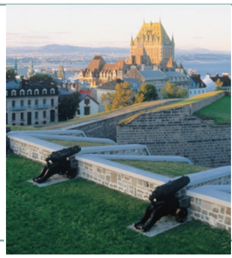
- "in statistics the problems always evolve relative to the development of new data structures and new computational tools" … rapport du NSF
- "Statistics is driven by data" … Don McLeish
- "Our discipline needs collaborations" … Hugh Chipman
- Comment créer des opportunités?
  - Comment établir une identité qui nous est propre?
- Face à des pressions bureaucratiques de fusionner?
- Continuer à mettre l'accent sur ce que nous faisons de meilleur!!



Spéculations

- Engle
  - Variabilité, modélisation, données, théorie, données, théorie
- Tibshirani
  - Validation croisée; “forensic statistics” (Baggerly & Coombes, AAOS, 2009 # 4)
- Netflix Grand Prize
  - “Recommender systems”: apprentissage machine, psychologie, statistique!
- Tufte
  - “Visual Display of Quantitative Information” -- 1983


Completed
More [+]
Map View:
Funding
Diversity
Unemployment
787,000,000,000 $
$0



Merci!!

Félix Labrecque

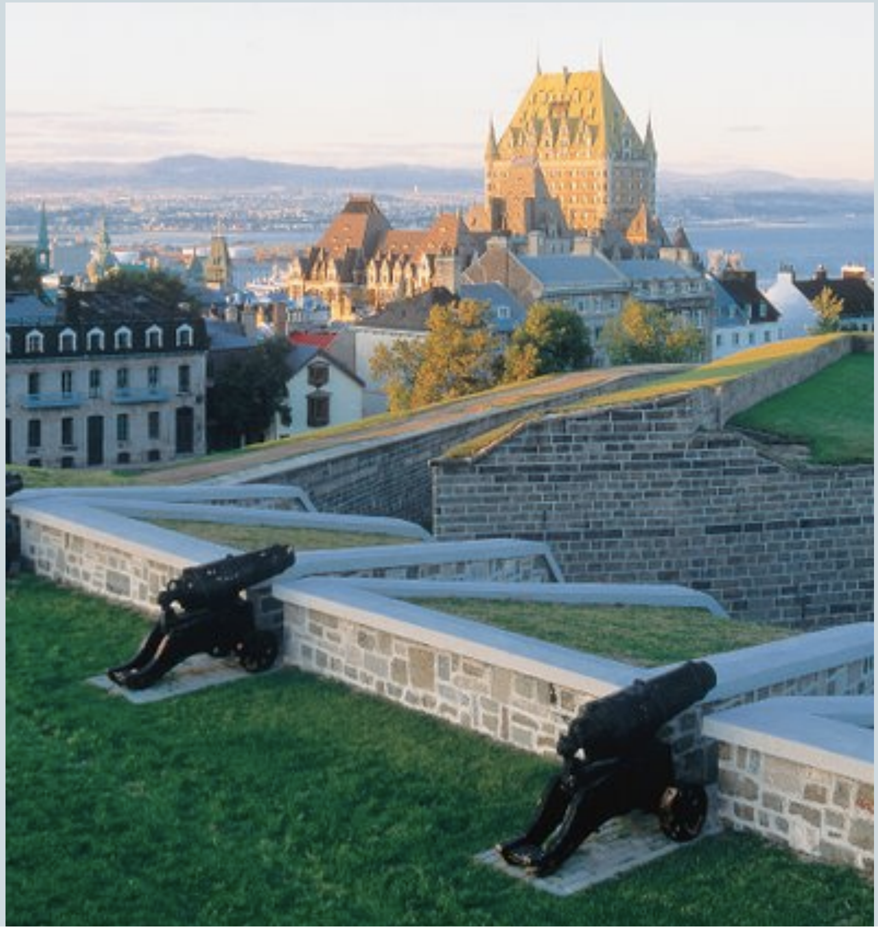
Cynthia Bocci

Jean-François Plante

Congrès Annuel 2010 à Québec

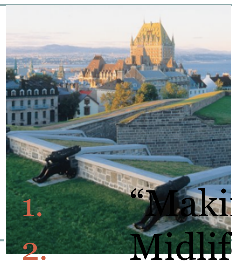
38e Congrès annuel de la Société statistique du Canada

Merci!!



Congrès Annuel 2010 à Québec

38e Congrès annuel de la Société statistique du Canada



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