

Stochastiek 2

Inleiding in de Mathematische Statistiek

Credible sets

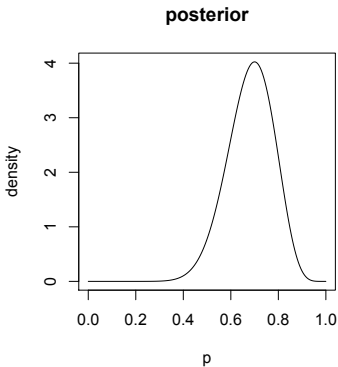
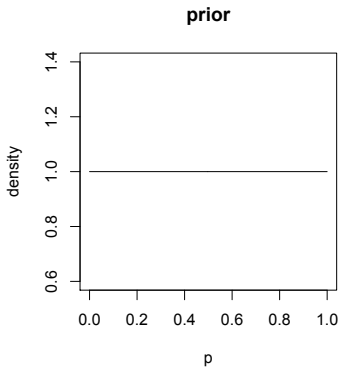
(Bayesiaanse “betrouwbaarheidsgebieden”)

Credible sets - 1

Bayesiaanse betrouwbaarheidsgebieden: **Credible sets**.

Data: $X \sim \text{Bin}(n, p)$, **prior** $\pi \sim \text{Beta}(\alpha, \beta)$.

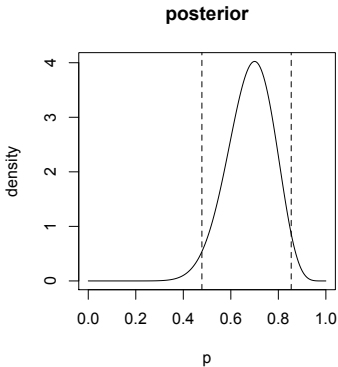
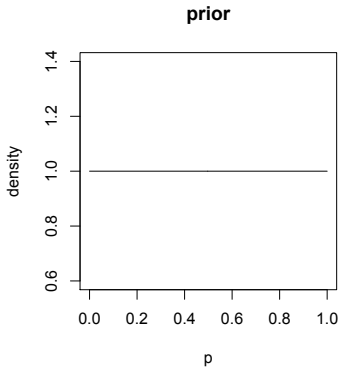
Posterior: $\text{Beta}(X + \alpha, n - X + \beta)$.



Credible sets - 2

Kies $\alpha \in (0, 1)$ (bijv. 0.05).

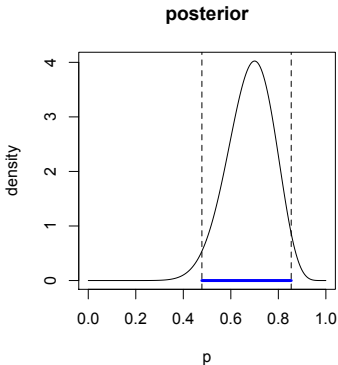
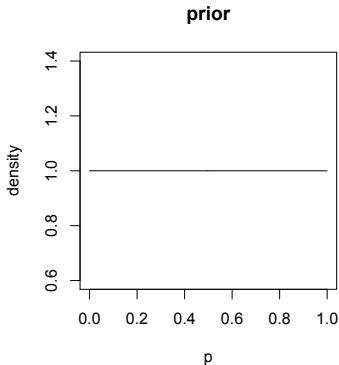
Bepaal gebied waar $1 - \alpha$ a-posteriori massa ligt:



Credible sets - 3

Dat gebied heet dan een $1 - \alpha$ **credible interval**.

Hier het **blauwe** interval:



Credible sets versus betrouwbaarheidsgebieden

- Een credible set is i.h.a. **niet** automatisch ook een betrouwbaarheidsgebied.
- In gladde, parametrische modellen is dat **asymptotisch** wel zo.
- Gevolg van **Bernstein-Von Mises**.

Credible sets versus betrouwbaarheidsgebieden

Zij X_1, X_2, \dots, X_n o.o. $\sim p_\theta$, voor $\theta \in \mathbb{R}$. "Regulier" model.

Def. $\ell_\theta(x) = \log p_\theta(x)$, $\dot{\ell}_\theta(x) = \frac{\partial}{\partial \theta} \ell_\theta(x)$ en $i_\theta = \text{Var}_\theta \dot{\ell}_\theta(X_1)$

BvM handout:

$$\int_B p_{\bar{\theta} | X_1, \dots, X_n}(\theta) d\theta \approx N\left(\theta_0 + \frac{\Delta_n}{\sqrt{n}}, \frac{1}{ni_{\theta_0}}\right)(B),$$

met

$$\Delta_n = \frac{1}{i_{\theta_0} \sqrt{n}} \sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i).$$

Credible sets versus betrouwbaarheidsgebieden

Een (symmetrisch) $1 - \alpha$ credible interval is dan asymptotisch van de vorm

$$I = \left[\theta_0 + \frac{\Delta_n}{\sqrt{n}} + \frac{1}{\sqrt{ni_{\theta_0}}} \xi_{\alpha/2}, \theta_0 + \frac{\Delta_n}{\sqrt{n}} + \frac{1}{\sqrt{ni_{\theta_0}}} \xi_{1-\alpha/2} \right].$$

We hebben

$$P_{\theta_0}(\theta_0 \in I) = P_{\theta_0}(\xi_{\alpha/2} \leq -\sqrt{i_{\theta_0}} \Delta_n \leq \xi_{1-\alpha/2}) \rightarrow 1 - \alpha,$$

omdat $-\sqrt{i_{\theta_0}} \Delta_n \xrightarrow{d} N(0, 1)$.

Dus in gladde, reguliere modellen: **een $1 - \alpha$ credible interval is een benaderd $1 - \alpha$ betrouwbaarheidsinterval.**

En nu ...

Wat hebben we gedaan

Setting: meestal X_1, \dots, X_n o.o. $X_i \sim p_\theta$, $\theta \in \Theta \subset \mathbb{R}^k$, $k = 1, 2$.

- Grafische methoden voor model check
- Methoden voor schatten van θ (ad hoc, MLE, Bayes, momenten)
- Methoden voor toetsen van hypothesen (Gauss-toets, binomiale toets, t-toetsen, likelihood-ratiotoets)
- Methoden voor betrouwbaarheidsgebieden ((bijna) pivots, likelihood-ratiogebieden, credible sets)
- Optimaliteitstheorie, basis asymptotiek voor MLS en Bayes

Wat is er nog meer te doen

- Grondslagen: **maattheoretische** en waarschijnlijkheidstheoretische achtergrond
- (Asymptotische) statistische **theorie**
- **Hoog-dimensionale/niet-parametrische statistiek**: schatten hoog/oneindig-dim parameter
- Statistisch **modelleren**: (gegeneraliseerde) lineaire modellen, ...
- Afhankelijke data: **tijdreeksen**.
- **Numerieke** methoden: EM, MCMC, bootstrap, ...
- **Toegepaste** statistiek: forensische statistiek, biostatistiek, netwerkstatistiek, ...
- **Machine Learning**

Vakken die je kan doen

Bachelor:

- Bayesiaanse statistiek (of Mathematische Statistiek)
- Statistische data analyse (VU)
- Forensische statistiek
- Tweede- en derdejaars projecten
- Maattheorie
- Markov ketens
- Analyse, Functionaalanalyse (lineaire analyse?)
- ...

Vakken die je kan doen

Master:

- Measure theoretic probability (mastermath)
- Asymptotic statistics (mastermath)
- Nieuw mastervak: **Statistical Theory for High- and Infinite-Dimensional Models**
- Time Series (mastermath)
- High-dimensional data analysis (VU)
- Simulation methods for statistics
- Statistics for Networks (VU)
- ...

Lineaire modellen

Voorbeeld

Data: $(\underbrace{Y_i}_{\text{response}}, \underbrace{x_{i,1}, \dots, x_{i,p}}_{\text{"verklarend"}}, i = 1, \dots, n.$

Model: $Y_i = \beta_1 x_{i,1} + \dots + \beta_p x_{i,p} + \varepsilon_i$, met ε_i i.i.d., verwachting 0.

Vragen:

- Schat de parameter $\beta = (\beta_1, \dots, \beta_p)$.
- Is β_i significant $\neq 0$?
- Welke verklarende variabelen moeten er in het model opgenomen worden?
- Wat als $p \gg n$?
- ...

Niet-parametrische statistiek

Voorbeelden:

- X_1, \dots, X_n o.o., met dichtheid f . Vraag: schat f .
- $(X_1, Y_1), \dots, (X_n, Y_n)$ o.o.,

$$Y_i = f(X_i) + \varepsilon_i,$$

met $\varepsilon_1, \dots, \varepsilon_n$ o.o., $N(0, \sigma^2)$. Vraag: schat f (en σ^2).

Vragen:

- Rate \sqrt{n} i.h.a. onmogelijk. Welke rate zijn wel mogelijk? (Hangt af van “complexiteit” van f , bijv gladheid!).
- Welke schatters halen optimale rates?

Statistical Learning

<http://mlg.eng.cam.ac.uk/pilco/>

Het wordt alleen maar leuker . . .