

Distribuzioni a priori

Strategie per la specificazione

Docente: Matilde Trevisani

DEAMS

A.A. 2024/2025
(aggiornato: 2024-10-28)

Agenda

Strategie nello specificare la a priori

- Specificare le a priori
 - Specificare una famiglia di distribuzioni
 - Specificare un elemento della famiglia
- Disegnare le a priori e la predittiva a priori
- Scambiabilità
- Non sentirsi a disagio ...
- Analisi di sensitività

Distribuzioni a priori noninformative (e debolmente informative)

- La a priori di Jeffrey

Strategie per specificare distribuzioni a priori

Specificare le a priori

- Step 1: Specificare una famiglia di distribuzioni
 - Es: N
- Step 2: Specificare un elemento della famiglia
 - Specificare gli **iperparametri**
 - Es: $N(0, 1)$

Specificare una famiglia di distribuzioni

Criteri per specificare una famiglia di distribuzioni

- Coerenza con conoscenze sostanziali
- Flessibilità/ricchezza
- Facilità di calcolo
- Facilità di interpretazione
- Ricerca metodologica

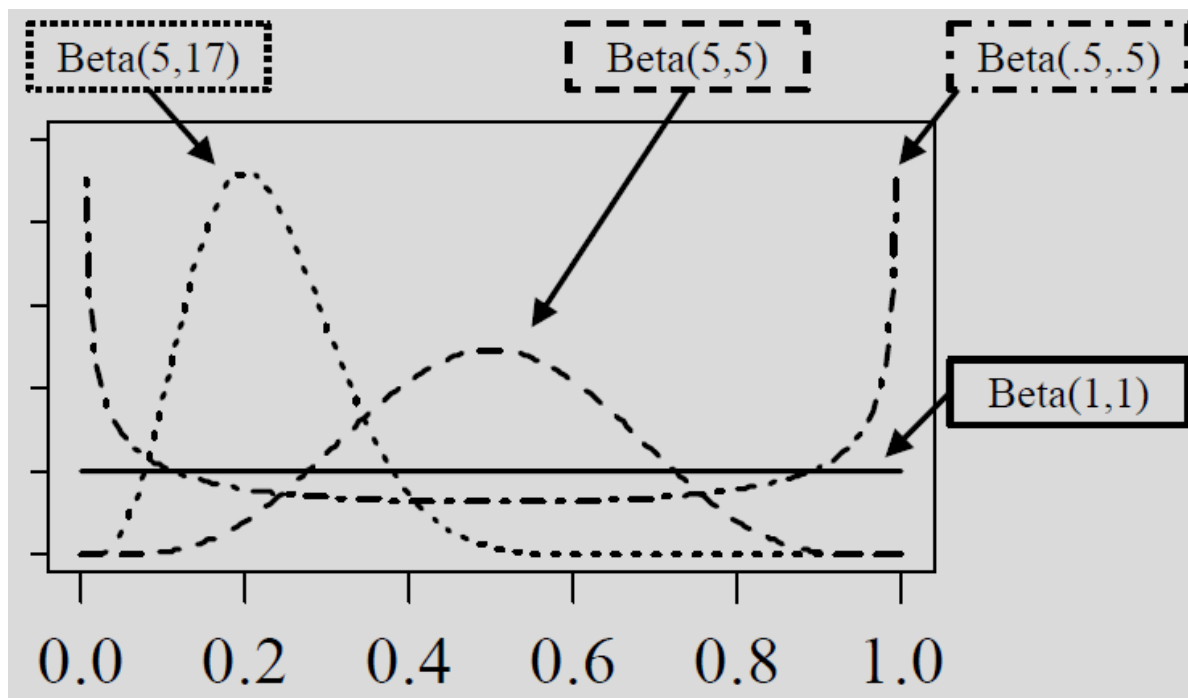
Coerenza con conoscenze sostanziali

- Specifica una a priori che rifletta le tue convinzioni
 - Quali possibilità (anche teoria, analisi precedenti)
- Almeno considera il campo di definizione:
 - Parametro $\in (-\infty, \infty)$: normale, t
 - E.g., costante di regressione, pendenza
 - Parametro > 0 : lognormale, gamma, gamma inversa, normale troncata
 - E.g., varianze, precisioni, coefficiente strutturale che sai essere > 0
 - Parametro $\in (0, 1)$: beta
 - E.g., proporzione, coefficiente strutturale standardizzato (che sai essere in $(0, 1)$)

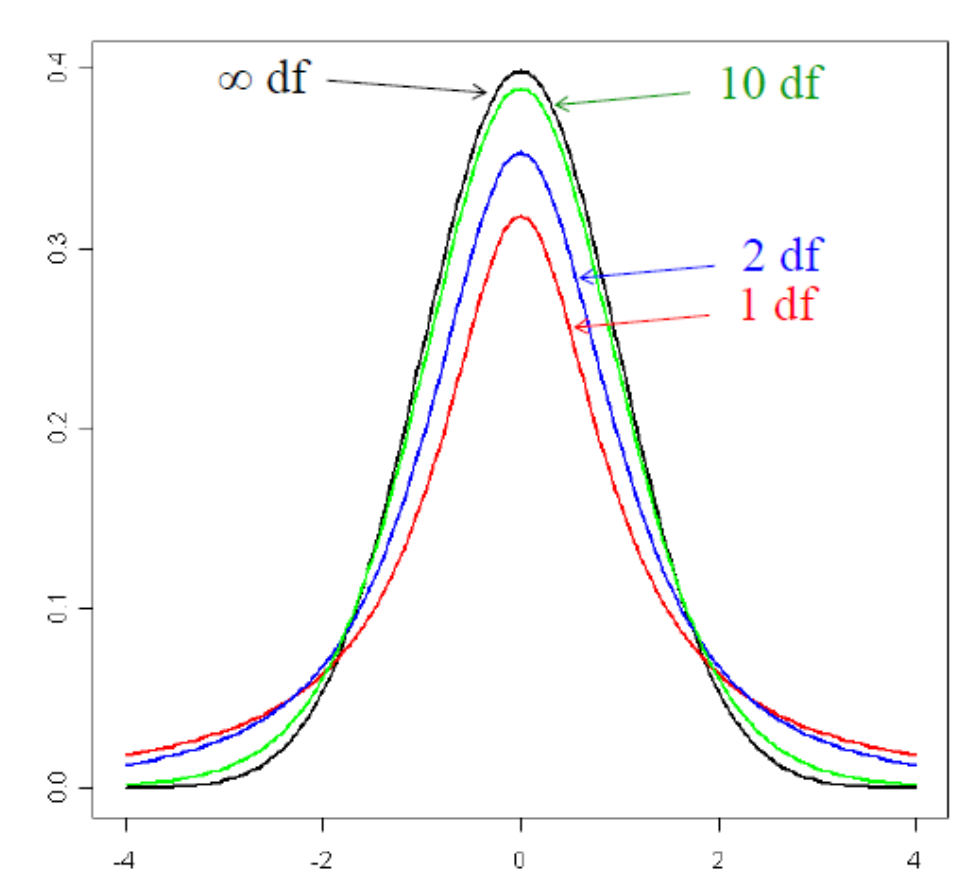
Flessibilità/Ricchezza

- Per un parametro $\in (0, 1)$, come una proporzione, si potrebbe usare una uniforme... o una beta, che sussume l'uniforme
- Per un parametro $\in (-\infty, \infty)$, come un coefficiente di regressione, si potrebbe usare una normale... o una t , che sussume la normale

Distribuzioni Beta



Distribuzioni t



Facilità di calcolo

- Con quanta facilità si ottiene la distribuzione a posteriori?
- *Coniugacy*: l'a priori è di una famiglia tale che l'a posteriori è di una famiglia nota (la stessa)
 - Si può ottenere l'a posteriori analiticamente

$$p(x|\theta) = \text{Bern o Bin}, p(\theta) = \text{Beta} \Rightarrow p(\theta|x) = \text{Beta}$$

$$p(x|\theta) = N, \sigma^2 \text{ noto}, p(\theta) = N \Rightarrow p(\theta|x) = N$$

Come trovare la coniugata?

- Esprimi la distribuzione condizionale dei dati in termini di statistiche sufficienti
- Considera queste come costanti che governano la distribuzione del parametro

Es. Binomiale: le statistiche sufficienti sono # successi (y) e # tentativi (J)

$$p(y|\theta) \propto \theta^y (1 - \theta)^{J-y}$$

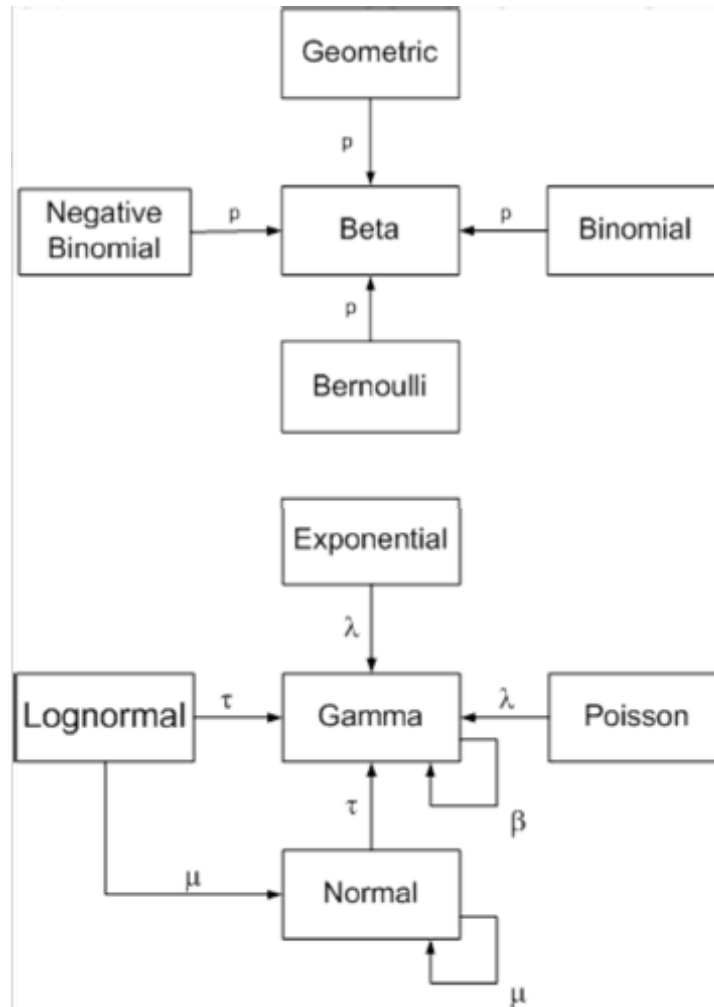
Se trattiamo y e J come costanti, e θ come una variabile

$$\theta^{const} (1 - \theta)^{const}$$

riconosciamo una Beta

Mappa delle distribuzioni coniugate

http://www.johndcook.com/conjugate_prior_diagram.html



Coniugata condizionalmente, semi- o generalizzata

- Nozione di coniugata condizionale, semi o generalizzata
 - Distribuzione a posteriori non di forma nota, ma può risultare computaz. trattabile, ad esempio, in alcuni algoritmi MCMC

$$p(x|\theta, \sigma^2) = N$$

Coniugata condizionalmente:

$$p(\theta, \sigma^2) = p(\sigma^2) p(\theta) = \text{Inv-Gamma } N$$

■ Lo vedremo con i modelli multiparametrici

Facilità di calcolo

- Con quanta facilità si ottiene la distribuzione a posteriori?
- *Coniugacy* > *coniugacy* condizionale > non coniugata
- In molti casi, non esiste un'a priori coniugata e potrebbe non esserci una coniugata generalizzata
- Meno preoccupazioni grazie a MCMC

Facilità di interpretazione

- L'a priori può essere interpretata in termini sostanziali?
- Di solito tramite gli iperparametri

A priori Beta per Binomiale

$$\begin{aligned} p(\theta|y, J) &\propto p(y|\theta, J) p(\theta) \\ &= \text{Bin}(y|\theta, J) \text{Beta}(\alpha, \beta) \\ &\propto \theta^y (1 - \theta)^{J-y} \theta^{\alpha-1} (1 - \theta)^{\beta-1} \end{aligned}$$

Suggerisce un'interpretazione per la a priori come
 $\alpha - 1$ successi e $\beta - 1$ fallimenti

Ricerca metodologica

- Esistono alcune ricerche sull'utilità di usare a priori differenti
- Campo di ricerca in continua espansione su come rappresentare un'a priori diffusa per la variabilità
 - Considera diverse a priori per varianza, precisione, deviazione standard
 - Inv-Gamma(α, β) per la varianza, equivalentemente Gamma(α, β) per la precisione, con $\alpha = \beta \approx 0$ (eg., .001), usate perchè sostenuto garantissero un'adeguata diffusione a priori
 - La ricerca metodologica ha contestato l'uso delle Gamma diffuse
 - Si suggeriscono altre distribuzioni come esponenziale, half-normal, half-Cauchy, uniforme e altre

Priorità nella definizione di una a priori

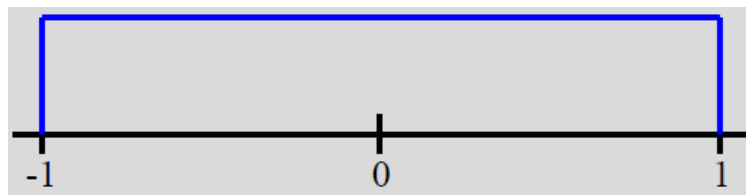
- Storicamente, le a priori sono stati scelte il più possibile diffuse
- Poi c'è stato un passaggio alla fattibilità computazionale (*conjugacy*)
 - Se non riesci a capire l'a posteriori, l'inferenza di Bayes non è così utile
 - Meno rilevante con approssimazioni (Bayes modale)
 - Molto meno rilevante con MCMC
- Ora, si consiglia di enfatizzare la coerenza con le conoscenze sostanziali
 - Si paga volentieri il prezzo del calcolo per modellare le proprie convinzioni
 - Flessibilità, facilità di interpretazione, tutte di supporto in quanto rendono più facile rappresentare le conoscenze sostanziali
- Il futuro trarrà vantaggio dalla ricerca metodologica

Specificare un elemento della famiglia

Specificare un elemento della famiglia

Strategia 1: quanta più ignoranza possibile

- Esempi
 - Normale con un'enorme varianza come a priori per la pendenza di una regressione
 - $U[-1, 1]$ a priori per la correlazione tra x e y



- Il principio è essere cauti nell'"iniettare" informazioni nell'a priori
- Ciò che conta per essere "non informativo" dipende dalla quantità dei dati, dalla scala delle variabili
 - e.g., $N(0, 10)$ diffusa per θ tra $(0, 1)$, non deve esserlo necessariamente altrove

Specificare un elemento della famiglia

Strategia 2: utilizzare dati, risultati precedenti

- Se 19.582 studi hanno suggerito che la correlazione tra x e y è compresa tra .4 e .6 si utilizzi un'a priori che metta la maggior parte della densità in quella regione (invece che un'a priori $U(-1, 1)$!)
- Conosci la tua area di ricerca!
- Meta-analisi

Specificare un elemento della famiglia

Strategia 3: Usa la teoria, la conoscenza dei SMEs

Usa la teoria

- E.g., per la correlazione, *credi davvero* che
 - un valore di .9 sia probabile quanto .4?; un valore di .9 abbia la stessa probabilità di 0?; un valore di .9 sia altrettanto probabile quanto -.9?
 - ... o credi che alcuni valori siano più o meno probabili?
- Esempio dalla misurazione nell'istruzione: p (indovinare la risposta corretta) su un elemento con:
 - 2 categorie; 4 categorie; 100 categorie, ...
 - probabile che sia non lontana da una probabilità uguale per item

Spesso la conoscenza proviene da esperti in materia (SMEs)

- Letteratura sulla elicitazione delle a priori
 - E.g., O'Hagan et al. (2006), von Winterfeldt e Edwards (1986), Winkler (1972), Stampa (1989), Kadane & Wolfson (1998), O'Hagan (1998)
- Esempi pratici:
 - “A priori linguistiche” di Almond (2010) nella teoria della risposta agli item, Reti di Bayes
- Software
 - Novick (1971) Software CADA per una varietà di modelli di base
 - Morris et al. (2014)
 - Strumento MATCH basato sul web di Uanhero (2021)
 - Altri che sono specifici del modello o della situazione

Specificare un elemento della famiglia

Strategia 4: teoria, dati, un pizzico di ignoranza

Ad esempio, Modello p (indovinare la risposta corretta) su un elemento con 5 opzioni

- Teoria: con 5 alternative, la possibilità di indovinare la risposta corretta è probabilmente circa .2 (ed è più probabile che sia vicino a .2 che a .9)
- Dati: se per gli elementi per i quali possiamo stimare la probabilità di indovinare la risposta corretta essa è da qualche parte tra .1 e .4, allora per il resto degli elementi, la possibilità di indovinare la risposta corretta è probabilmente compresa tra .1 e .4
- Ignoranza: voglio usare un'a priori che rifletta queste convinzioni, ma non voglio darci troppo peso
 - E.g., $Beta(5, 17) \rightarrow$ come se avessi visto 4 successi in 20 prove $\rightarrow$ vd. Beta's

Disegnare le a priori e la predittiva a priori

Vd. file *Materiale ...*, *Example*, p.28

Scambiabilità

Vd. esempio p.41

**Spazzare via quella sensazione di disagio riguardo alle
a priori**

Inquietudine nello specificare a priori

- Spesso si esita più sull'opportunità di utilizzare a priori, piuttosto che come
- Consideriamo i motivi per cui si potrebbe obiettare a farlo

Obiezione: non avere convinzioni a priori

■ "Non ho convinzioni a priori."

- Se vedi un arciere colpire il bersaglio 6 volte su 6, ***pensi davvero*** che il suo $p(hit) = 1$?
- Per la correlazione (pendenza standardizzata, ecc.), ***credi davvero*** che un valore di .9 è altrettanto probabile di -.9?
- Quando una stima di una varianza risulta < 0 , ci credi davvero?
- ... Percentuale di tali pazienti con cancro al seno = .008
- ... Fisica
- ... Psicologia
- Non hai convinzioni a priori $\Rightarrow$ non hai fatto i compiti

Obiezione: Non vogliamo soggettività

“Posso avere convinzioni a priori, ma non voglio che esse offuschino l'analisi. L'analisi dovrebbe essere oggettiva.”

Ad esempio, le applicazioni tipiche della regressione assumono $y|x \sim N$

- La verosimiglianza assume la forma di una N
- Perché usare una N , invece di t per la verosimiglianza?
- Perché specificare nella verosimiglianza un modello lineare? Con quei parametri, e non altri?
- Più in generale, *perché hai usato questa verosimiglianza?*
- Come determini gli outliers? Come decidi cosa fare di loro?
- La tua scelta di α ? n ? $\mathbf{X}$?
- Quindi, molte scelte sono soggettive
 - vd. sopra
 - ...
- Soggettivo $\neq$ Arbitrario

Analisi di sensitività

Analisi di sensitività

- Condurre l'analisi utilizzando più a priori
- Se le inferenze, conclusioni, decisioni, ecc., a posteriori, sono le stesse, ciò suggerisce che le inferenze sono robuste rispetto alla scelta delle a priori (vd. p. 50)
- Se le inferenze, conclusioni, decisioni, ecc., a posteriori, differiscono, ciò conduce a un'ulteriore discussione e considerazione della scelta delle a priori (vd. p. 52)
- Non viene fatta spesso, ma alcuni dicono che dovrebbe esserlo (cfr. verosimiglianza)
- Altri dicono che non è così importante

“La mia a priori... è la mia a priori. Se avessi una convinzione a priori diversa, utilizzerei una a priori diversa. Se hai in mente un'altra a priori, vai avanti dritto e usala per la tua inferenza.”

-->

Riepilogo

- Specificare le a priori
 - Coerenza con le conoscenze sostanziali
 - Flessibilità/ricchezza
 - Facilità di calcolo
 - Facilità di interpretazione
 - Ricerca metodologica
- Utilizzo di dati esistenti, teoria
- Scambiabilità
- Sensitività
- Allargare gli orizzonti
- Area di ricerca attiva

Esempi (pag 57)

- A priori basata su ricerca precedente
- A priori basata su opinione degli esperti
- Es su *American College Testing* (a standardized test used for college admissions in the United States)

Distribuzioni a priori non-informative

Distribuzioni a priori non-informative

- *Reference prior*: "Let the data speak for themselves"

Reference analysis produces objective Bayesian inference, in the sense that inferential statements depend only on the assumed model and the available data, and the prior distribution used to make an inference is least informative in a certain information-theoretic sense. (Berger et al, 2009)

- Molte volte si usano a priori improprie come *reference priors*
- Es: media non nota, var nota

$$y|\theta \sim N(\theta, \sigma^2), \theta \sim N(\mu_0, \tau_0^2) \Rightarrow \theta|y \sim N(\theta|\mu_1, \tau_1^2)$$

Se $\tau_0^2 = \infty$ e n fissato, (o $n = \infty$ e τ_0^2 fissato)

$$p(\theta) \propto \text{constant} \Rightarrow \theta|y \approx N(\theta|\bar{y}, \sigma^2/n)$$

Si noti che $p(\theta) \propto \text{constant}$, $\theta \in (-\infty, \infty)$, è impropria

NB: A priori improprie possono portare ad a posteriori proprie

- Es: media nota, var non nota

$$y|\sigma^2 \sim N(\theta, \sigma^2)$$

$$\sigma^2 \sim IG(\alpha, \beta)$$

Se $\alpha = \beta = 0$, ovvero $\nu_0 = 0$ per la chi-quadro inversa scalata equivalente, allora

$$p(\sigma^2) \propto \frac{1}{\sigma^2}$$

$$\Rightarrow \sigma^2|y \approx \text{Inv-}\chi^2(n, v = \sum_i (y_i - \mu)^2/n)$$

si noti che $p(\sigma^2) \propto \frac{1}{\sigma^2}$ è impropria avendo integrale infinito nel range $(0, \infty)$

NB: A priori improprie possono portare ad a posteriori proprie

Principio di invarianza di Jeffreys

Un approccio che a volte viene utilizzato per definire distribuzioni a priori non informative venne introdotto da Jeffreys ed è basato sul considerare delle trasformazioni uno-a-uno del parametro: $\phi = h(\theta)$.

Il principio generale di Jeffreys è che qualsiasi **regola per determinare la densità a priori** $p(\theta)$ dovrebbe comportare un **risultato equivalente** se applicato al parametro trasformato ϕ ; cioè, $p(\phi)$ calcolato determinando prima $p(\theta)$ e applicando poi il **teorema del cambio di variabile**, i.e.,

$$p_{\phi}(\phi) = p_{\theta}(h^{-1}(\phi)) |J| = p(\theta) \left| \frac{d\theta}{d\phi} \right|$$

dovrebbe corrispondere alla distribuzione $p(\phi)$ ottenuta utilizzando la "regola" direttamente sul modello trasformato, $p(y, \phi) = p(\phi)p(y|\phi)$.

N.B. Nel caso multivariato $|J|$ è il determinante dello Jacobiano J della trasformazione inversa $\theta = h^{-1}(\phi)$, dove J è la matrice (quadrata) delle derivate parziali con elemento (i, j) uguale a $\partial\theta_i/\partial\phi_j$.

Definizione della a priori di Jeffreys

Il principio di Jeffreys porta a definire la densità a priori non informativa come

$$p(\theta) \propto [I(\theta)]^{1/2}$$

dove $I(\theta)$ è l'*informazione di Fisher* per θ :

$$I(\theta) = E \left(\left(\frac{d \log p(y|\theta)}{d\theta} \right)^2 | \theta \right) = -E \left(\frac{d^2 \log p(y|\theta)}{d\theta^2} | \theta \right)$$

Infatti, valutando $I(\phi)$ in $\theta = h^{-1}(\phi)$, si dimostra che

$$I(\phi) = I(\theta) \left| \frac{d\theta}{d\phi} \right|^2$$

da cui l'invarianza è soddisfatta.

Varie a priori non-informative per il parametro della Binomiale

- $y|\theta \sim \text{Bin}(y|\theta, n)$
- $I(\theta) = \frac{n}{\theta(1-\theta)}$ da cui la a priori di Jeffrey

$$p(\theta) \propto \theta^{-1/2}(1 - \theta)^{-1/2} \propto \text{Beta}(1/2, 1/2)$$

- A priori Bayes-Laplace è l'uniforme

$$p(\theta) \sim \text{Beta}(1, 1)$$

- A priori uniforme per il $\text{logit}(\theta)$

$$p(\text{logit}(\theta)) \propto \text{constant}$$

corrisponde ad una $\text{Beta}(0, 0)$ per θ , ed è impropria

ATTENZIONE: se l'a priori è $\text{Beta}(0, 0)$ e $y = 0$ o $y = n$ allora l'a posteriori è impropria!

Note su a priori non-informative

- Se la verosimiglianza è veramente dominante, allora la scelta nella gamma di a priori relativamente piatte è sostanzialmente indifferente
- Una densità piatta per una data parametrizzazione può non essere piatta in un'altra, e.g., $p(\log \sigma^2) = 1$ ma $p(\sigma^2) = 1/\sigma^2$
- Vantaggio: non sembra valga la pena esprimere la reale conoscenza a priori come una distr di prob propria se si controlla che la distr a post sia propria e si effettui l'analisi di sensitività