Statistica Bayesiana

Pagani Davide

11 novembre 2016

Indice

1 Introduzione

Inferenza classica (o frequentista, nata nel '20 con Fisher): ε è un esperimento aleatorio o casuale di cui non si conosce il risultato, facendo n prove indipendenti ottengo $\underline{x} = (x_1, x_2, ..., x_n)$, una $n - upla$ campionaria che contiene **l'informazione campionaria**, alla quale si aggiunge il principio del campionamento ripetuto che mi permette di passare dalla stima di un numero allo stimatore $\underline{x} \to \underline{X}$.
L'inferenza bayesiana invece è più recente (anni '50, '60 con Bayes): si utilizza ancora l'informazione campionaria $\underline{x}$ alla quale si aggiunge l'informazione pre-sperimentale, cioè ho qualcosa in più prima di fare l'esperimento (uso il principio di verosimiglianza).

Esempio (numero guasti degli impianti produttivi)
Per analizzare il numero di guasti in un impianto produttivo utilizzo una variabile casuale di Poisson con parametro θ (ignoto).
$X \sim Poisson(\theta) \to$ voglio fare inferenza su θ quindi sul parametro ignoto e non aleatorio che corrisponde al numero dei guasti.
In ambito classico avrei calcolato la stima di massima verosimiglianza che coincide con la media campionaria: $\hat{\theta} = \frac{\sum x_i}{n}$; **In ambito bayesiano** ho delle informazioni in più (ad esempio so che $\theta < 20$ oppure so che è più probabile un numero piccolo di guasti rispetto ad uno elevato) che devo integrare nel mio problema. Quindi non ho più θ ignoto.

Vedo dal grafico che in un intervallo di sinistra ho una probabilità più elevata rispetto a destra + informazione pre-sperimentale all'interno della mia statistica + principio di verosimiglianza (tutto ciò che è sintetizzato nell'$n-upla$ campionaria presa in considerazione dato che mi da tutta l'informazione di cui ho bisogno).

Notazione simbolica:
Modello (esperimento) statistico: $(X, f(\underline{x}; \theta), \theta \in \Theta)$.

Esempio (fenomeno dicotomico)
Fenomeno dicotomico:

$$(\{0, 1\}, f(x; \theta) = \theta^x (1 - \theta)^{1-x}, \theta \in [0, 1])$$

Con 0 = insuccesso e 1 = successo.

Dopo aver fatto n prove indipendenti il modello indotto è:

$$(X^{(n)}, f(\underline{x};\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}, \theta \in [0,1])$$

Con $X^{(n)}$ che corrisponde a tutte le $n-uple$ in cui ciascun elemento si muove in x, cioè in un fenomeno dicotomico.
$f(\underline{x};\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$ corrisponde alla funzione di verosimiglianza $L(\theta)$.

Esempio (Savage '62)
Il campione è $\underline{x} = (1, 1, ..., 1)$ con $n = 10$, quindi la stima di massima verosimiglianza è la media campionaria $\hat{\theta} = \frac{\sum x_i}{n} = \overline{x} = \frac{10}{10} = 1$; è molto verosimile pensare che il parametro ignoto sia 1. Savage dice che prima dell'esperimento ognuno ha una propria idea su quello che può essere il valore di $\theta \rightarrow$ prima di fare l'esperimento, devi avere in testa dei valori plausibili; ognuno avrà la propria idea in base all'esperienza pregressa. Se non so, avrò una funzione con area al di sotto sempre uguale.

Per aggiungere l'informazione pre-sperimentale al modello passo dalla funzione di verosimiglianza alla legge condizionata $f(\underline{x}|\theta)$.
La legge della variabile θ è detta legge a priori $\pi(\theta)$, dalla quale si può ricavare la legge di distribuzione congiunta:

$$\Psi(\underline{x}, \theta) = f(\underline{x}|\theta)\pi(\theta)$$

$f(\underline{x}|\theta)$ è la legge condizionata al valore di θ manifestata nel momento in cui ho fatto l'esperimento.
$\pi(\theta)$ invece è la prior $\rightarrow$ legge di distribuzione marginale.

Per calcolare la posterior uso la congiunta:

$$\pi(\theta|\underline{x}) = \frac{\Psi(\underline{x}, \theta)}{m(\underline{x})} = \frac{f(\underline{x}|\theta)\pi(\theta)}{\int_\theta f(\underline{x}|\theta)\pi(\theta)\, d\theta} = c \cdot f(\underline{x}|\theta)\pi(\theta)$$

dove c è la **costante di normalizzazione** $\frac{1}{m(\underline{x})}$ (reciproco della marginale di x) mentre il resto del prodotto è il **nucleo** della posterior.

$$\textbf{Teorema di Bayes} : \frac{f(\underline{x}|\theta)\pi(\theta)}{\int_\theta f(\underline{x}|\theta)\pi(\theta)\, d\theta}$$

Esempio (Gamma)

$$X \sim Gamma(\alpha, \beta)$$

con $\alpha, \beta > 0$, e $x > 0$

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} exp(-\beta x)$$

la parte a sinistra $\frac{\beta^\alpha}{\Gamma(\alpha)}$ è la costante di normalizzazione che non dipende da x e permette all'integrale della densità di essere pari a 1, la parte a destra $x^{\alpha-1}exp(-\beta x)$ è il nucleo della legge di distribuzione.

Esempio (Bernoulli)
Data la variabile $X \sim Bernoulli(\theta)$ vengono fatte 3 ($n = 3$) prove e il campione ottenuto è $\underline{x} = (0, 0, 1)$.
La prior $\pi(\theta) = 1$
Modello statistico indotto:

$$(\{0,1\}^{(3)}, \theta(1-\theta)^2, \theta \in [0,1])$$

$$\pi(\theta|\underline{x}) = \frac{f(\underline{x}|\theta)\pi(\theta)}{m(\underline{x})} = \frac{\theta(1-\theta)^2 \cdot 1}{\int_0^1 \theta(1-\theta)^2 \cdot 1 d\theta}$$

$\theta(1-\theta)^2 \cdot 1 \rightarrow$ **nucleo**

$\int_0^1 \theta(1-\theta)^2 \cdot 1 d\theta \rightarrow$ **costante di normalizzazione.**

1.1 Variabile casuale Beta

$X \sim Beta(\alpha, \beta)$ ha supporto $[0, 1]$ o $(0, 1)$ mentre $\alpha, \beta > 0$.

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

Il nucleo integra a 1 solo con $\alpha = 1$ e $\beta = 1$ che sarebbe **uniforme continua**.

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx$$

$E[X] = \frac{\alpha}{\alpha+\beta}$

$Moda[X] = \frac{\alpha-1}{\alpha+\beta-2} \rightarrow$ solo se $\alpha, \beta > 1$ (altrimenti non c'è moda poichè sarebbe uniforme)

$Var[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

2 Modelli

2.1 Modello Binomiale-Beta

Considerando n prove indipendenti con fenomeno dicotomico:
$\{0,1\} \to \{insuccesso, successo\}$

$X \sim Bernoulli(\theta)$

Otteniamo una n-upla campionaria, una serie di $0, 1$.

Se $X \sim Beta(\alpha, \beta)$, allora è continua sul supporto $\{0,1\}$ con:

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$

dove $x^{\alpha-1}(1-x)^{\beta-1} \to$ **nucleo** e $B(\alpha,\beta) \to$ **costante di normalizzazione** che integra a 1, cioè $\int = 1$.

Considerando quindi come **prior** $\pi(\theta)$ di tipo $Beta(\alpha,\beta)$ tale che:

$$\pi(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \quad con\ 0 < \theta < 1\ e\ \alpha,\beta > 0$$

La fase di scelta della prior si chiama **elicitazione**
I parametri che compongono la prior vengono definiti **iperparametri** e compaiono nella legge di distribuzione di θ.

Invece la **posterior** è:
- Caso continuo:

$$\pi(\theta|\underline{x}) = \frac{f(\underline{x}|\theta)\pi(\theta)}{m(\underline{x})} = \frac{f(\underline{x}|\theta)\pi(\theta)}{\int_\Theta f(\underline{x}|\theta)\pi(\theta)\,d\theta}$$

- Caso discreto:

$$\pi(\theta|\underline{x}) = \frac{f(\underline{x}|\theta)\pi(\theta)}{m(\underline{x})} = \frac{f(\underline{x}|\theta)\pi(\theta)}{\sum_\Theta f(\underline{x}|\theta)\pi(\theta)\,d\theta}$$

Ricordo che $m(\underline{x})$ corrisponde alla marginale di $x \to$ **distribuzione predittiva iniziale**.

La funzione di verosimiglianza partendo da n prove indipendenti di una

$Bernoulli \rightarrow \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$, deve quindi essere moltiplicata per la prior di una $Beta \rightarrow \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$ in modo tale da ottenere:

$$\pi(\theta|\underline{x}) = \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \cdot \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}}{\int_0^1 \frac{\theta^{\sum x_i+\alpha-1}(1-\theta)^{n-\sum x_i+\beta-1}}{B(\alpha,\beta)} d\theta}$$

Il supporto $\Theta \in \{0,1\}$ utilizza l'integrale nel caso continuo e la sommatoria nel caso discreto.
Dalla formula della posterior è **sempre** possibile semplificare la costante $(B(\alpha,\beta))$

Possiamo distinguere il nucleo dalla costante di normalizzazione:

$$\theta^{\sum x_i+\alpha-1}(1-\theta)^{n-\sum x_i+\beta-1} \rightarrow \textbf{nucleo}$$

$$\int_0^1 \theta^{\sum x_i+\alpha-1}(1-\theta)^{n-\sum x_i+\beta-1} d\theta \rightarrow \textbf{costante di normalizzazione}$$

Il nucleo comprende i fattori della posterior che contengono θ.
La costante di normalizzazione considera tutti gli elementi che non contengono θ, poichè integrando per θ ottengo la marginale di x.

N.B. se riconosco il nucleo come variabile nota, la mia soluzione è il numeratore. Se non riconosco la distribuzione, passerò per via computazionale.

Dal nucleo osservo che nella funzione di densità avevo x, ora ho θ che eleva ad un numero. Gli iperparametri α e β della distribuzione $Beta$ li ricavo nel seguente modo:

- per α: prendo l'esponente di θ e lo sommo a $+1 \rightarrow \sum x_i+\alpha-1+1 = \sum x_i+\alpha$
- per β: prendo l'esponente di $(1-\theta)$ e lo sommo a $+1 \rightarrow n-\sum x_i+\beta-1+1 = n-\sum x_i+\beta$

$\pi(\theta|\underline{x}) = Beta(\sum x_i+\alpha, n-\sum x_i+\beta)$

Considerazione sul denominatore:

$$\int_0^1 \frac{\theta^{\sum x_i+\alpha-1}(1-\theta)^{n-\sum x_i+\beta-1}}{B(\sum x_i+\alpha, n-\sum x_i+\beta)} d\theta \cdot B(\sum x_i+\alpha, n-\sum x_i+\beta)$$

Quindi la posterior:

$$\pi(\theta|\underline{x}) = \frac{\theta^{\sum x_i+\alpha-1}(1-\theta)^{n-\sum x_i+\beta-1}}{B(\sum x_i+\alpha, n-\sum x_i+\beta)}$$

è la funzione di densità di una *Beta* con parametri α e β rispettivamente a $\sum x_i + \alpha$ e $n - \sum x_i + \beta$

Una via più **"breve"**:

$$\pi(\theta|\underline{x}) = \frac{f(\underline{x}|\theta)\pi(\theta)}{m(\underline{x})} \propto f(\underline{x}|\theta)\pi(\theta)$$

Posso quindi trascurare le costanti moltiplicative!

$\propto = \frac{1}{m(\underline{x})}$ significa "proporzionale a" e si utilizza quando si riesce a riconoscere il nucleo di una variabile casuale nota. Posso concludere che la distribuzione a posteriori coincide con la legge di questa variabile casuale nota.

Esempio (sottoscrizioni polizze giornaliere)
Voglio fare inferenza sul numero medio di polizze, cioè su θ.
$X \sim Poisson(\theta)$ con θ ignoto che corrisponde al numero medio di polizze $\rightarrow$ valore atteso della variabile aleatoria.
Considero 10 giorni, quindi una numerosità $n = 10$. Non è necessario sapere che $\underline{x} = (3, 10, 7, 4, ...)$ ma è importante sapere che $\sum_{i=1}^{10} x_i = 30$.
La stima di massima verosimiglianza di θ: $\hat{\theta} = \frac{\sum x_i}{n} = \frac{30}{10} = 3$ (corrisponde alla media campionaria)

Il modello indotto (terna di oggetti) è: $(\mathbb{N}^{(10)}, \frac{e^{-n\theta}.\theta^{\sum x_i}}{\prod_{i=1}^{n} x_i!}, \Theta = \mathbb{R}^+)$

Elicitazione prior: sappiamo a priori che la prior è una $Gamma(4, 1)$ tale che $\pi(\theta) = \Gamma(4, 1)$

$E(\theta) = \frac{\alpha}{\beta} = \frac{4}{1} = 4$
$VAR(\theta) = \frac{\alpha}{\beta^2} = \frac{4}{1^2} = 4$
$MODA(\theta) = \frac{\alpha-1}{\beta} = \frac{3}{1} = 3$ con $\alpha > \beta$

Generalizzando a n elementi, ci riconduciamo ad un **modello Poisson-Gamma**.

2.2 Modello Poisson-Gamma

$$\pi(\theta|\underline{x}) = \frac{f(\underline{x}|\theta)\pi(\theta)}{m(\underline{x})} = \frac{\frac{e^{n\theta}.\theta^{\sum x_i}}{\prod_{i=1}^{n} x_i!} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}}{\int_0^\infty \frac{e^{n\theta}.\theta^{\sum x_i}}{\prod_{i=1}^{n} x_i!} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}d\theta}$$

Raccolgo ciò che può essere raccolto nella posterior per poi eliminare tutto ciò che non dipende da θ come $\prod_{i=1}^{n} x_i!$, β^{α} e $\Gamma(\alpha)$:

$$\frac{\frac{\theta^{\sum x_i+\alpha-1}\cdot e^{-n\theta-\beta\theta}\cdot\beta^{\alpha}}{\prod_{i=1}^{n} x_i!\cdot\Gamma(\alpha)}}{\int_0^{\infty}\frac{\theta^{\sum x_i+\alpha-1}\cdot e^{-n\theta-\beta\theta}\cdot\beta^{\alpha}}{\prod_{i=1}^{n} x_i!\cdot\Gamma(\alpha)}d\theta} \longrightarrow \frac{\theta^{\sum x_i+\alpha-1}\cdot e^{-n\theta-\beta\theta}}{\int_0^{\infty}\theta^{\sum x_i+\alpha-1}\cdot e^{-n\theta-\beta\theta}d\theta}$$

Da cui ricavo $\Gamma(\sum x_i+\alpha, n+\beta)$ poichè ho una distribuzione nota. Considerando i risultati ottenuti dall'esempio delle sottoscrizioni polizze giornaliere, ottengo una $\Gamma(34, 11)$ poichè $\sum x_i = 30$, $\alpha = 4$, $n = 10$ e $\beta = 1$.

$E(\pi(\theta|\underline{x})) = \frac{34}{11} = 3,091$. Il numero medio di polizze giornaliere si è abbassato da 4 a $3,091$.
$VAR(\pi(\theta|\underline{x})) = \frac{34}{11^2} = 0,281$ da cui osserviamo che più il valore della varianza è piccolo e più siamo vicini al valore, quindi ho più fiducia.
$MODA(\pi(\theta|\underline{x})) = \frac{33}{11} = 3$ lo stesso valore della prior $\Gamma(4, 1)$.

Conclusione:
la $\Gamma(34, 11)$ risulta essere più precisa di una $\Gamma(4, 1)$ dato che è concentrata sui valori centrali della distribuzione.
Si può osservare inoltre che a parità tra 2 *Gamma*, più β è grande e più si ha maggior precisione e scarsa dispersione.

Considerazione:
Inferenza classica: $\hat{\theta} = \overline{x} = 3$
Prior: $E(\theta) = \frac{\alpha}{\beta}$
Posterior: $E(\theta|\underline{x}) = \frac{\sum x_i+\alpha}{n+\beta} = \frac{\frac{\sum x_i}{n}\cdot n+\frac{\alpha}{\beta}\cdot\beta}{n+\beta} = \frac{\hat{\theta}\cdot n+E(\theta)\cdot\beta}{n+\beta}$

2.3 Modello Esponenziale-Gamma

Il modello Esponenziale-Gamma viene utilizzato per fenomeni di durata, più in generale l'esponenziale negativa.

Considerando di aver rilevato la durata dei cellulari in anni:
$X \sim EsponenzialeNegativa(\theta)$ con funzione di densità $f(x;\theta) = \theta\cdot e^{-\theta x}$ sapendo che $x > 0$ e $\theta > 0$
$E(\theta) = \frac{1}{\theta}$, $n = 100$, $\sum_i x_i = 52,5$ anni e $SVM = \hat{\theta} = \frac{1}{\overline{x}}$

Formalizzando il modello statistico indotto: $(\mathbb{R}^{+(n)}, \theta^n\cdot e^{-\theta\sum x_i}, \Theta = \mathbb{R}^{+})$

Essendo in $\mathbb{R}^+$, provo ancora con la *Gamma* per la prior: $\pi(\theta) = \Gamma(\alpha, \beta)$. Propongo quindi una *Gamma* con $\alpha = 1$ e $\beta = 1$.

La posterior sarà quindi:

$$\pi(\theta|\underline{x}) \propto \theta^n \cdot e^{-\theta \sum_i x_i} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}$$

Eliminando i fattori che non coinvolgono θ

$$\propto \theta^{n+\alpha-1} \cdot e^{-\theta(\sum x_i + \beta)}$$

$\pi(\theta|\underline{x}) = \Gamma(n + \alpha, \sum_i x_i + \beta) = \Gamma(101, 53, 5)$
con $\alpha = 1$, $\beta = 1$, $n = 100$ e $\sum x_i = 52, 6$

Prior:
$VAR(\theta) = 1$
$MODA(\theta) = 1$
Posterior:
$VAR(\theta|\underline{x}) = 0, 035 \rightarrow$ avendo la varianza più piccola della prior, mi fido di più.
$MODA(\theta|\underline{x}) = 1, 86$

Considerazione:
Inferenza classica: $\hat{\theta} = \frac{1}{\underline{x}} = 1, 9 \rightarrow$ è un valore plausibile di θ che indica il tasso di rinnovo. Quindi vuol dire che su 100 persone che corrisponde al nostro campione, tendono a rinnovare il celluare 2 volte l'anno.
Prior: $E(\theta) = 1 \rightarrow$ rinnovo un cellulare all'anno
Posterior: $E(\theta|\underline{x}) = \frac{n+\alpha}{\sum x_i + \beta} = \frac{\frac{n}{\sum x_i} \cdot \sum x_i + \frac{\alpha}{\beta}\beta}{\sum x_i + \beta} = \frac{\hat{\theta} \sum x_i + E(\theta)\beta}{\sum x_i + \beta}$

2.4 Modello Normale-Normale

Considerando n prove indipendenti e identicamente distribuite in una normale tale che

$$X \sim N(\mu, \sigma^2)$$

con media μ ignota e varianza σ^2 nota.
Dato che sono in una normale con le precedenti notazioni per media e varianza, posso scrivere la prior come:

$$\pi(\mu) = N(\mu_0, \sigma_0^2)$$

Gli iperparametri μ_0 e σ_0^2 hanno valore numerico prefissato, quindi sono fissi.

Ciò che voglio fare è calcolare la posterior $\pi(\mu, \underline{x})$.

Per calcolare la posterior, abbiamo bisogno di:

1) Funzione di verosimiglianza:

$$L(\mu) = f(\underline{x}|\mu) = \frac{1}{(2\pi)^{\frac{n}{2}}} \cdot \frac{1}{(\sigma)^n} \cdot exp\{-\frac{1}{2}\sum(\frac{x_i - \mu}{\sigma})^2\}$$

Posso trascurare la prima parte della funzione di verosimiglianza poichè non dipende dl parametro μ.

$$\propto exp\{-\frac{1}{2\sigma^2}(\sum x_i^2 + n\mu^2 - 2\mu\sum x_i\}$$

Dato che la costante $(\sum x_i^2)$ non dipende da μ ed è nell'esponenziale, posso trascurarla.

$$\propto exp\{-\frac{1}{2\sigma^2}(n\mu^2 - 2\mu n\overline{x}\}$$

Ho riscritto $2\mu\sum x_i$ come $2\mu n\overline{x}$ dato che non cambia nulla.

2) Prior:

$$\frac{1}{\sqrt{(2\pi\sigma_0^2)}} \cdot exp\{-\frac{1}{2}(\frac{\mu - \mu_0}{\sigma_0})^2\}$$

Dato che è una legge di distribuzione per μ.
Posso trascurare ciò che non dipende da μ e ottengo

$$exp\{-\frac{1}{2\sigma_0^2}(\mu^2 - 2\mu_0\mu + \mu_0^2)\}$$

Ora posso trascurare μ_0^2 poichè è una costante moltiplicativa che non dipende da μ.

Dalla funzione di verosimiglianza e dalla prior otteniamo la posterior:

$$\pi(\mu|\underline{x}) \propto exp\{-\frac{1}{2\sigma^2}(n\mu^2 - 2\mu n\overline{x}\} \cdot exp\{-\frac{1}{2\sigma_0^2}(\mu^2 - 2\mu_0\mu)\}$$

Che ci darà come risultato

$$exp\{-\frac{1}{2(Var(\mu|\underline{x}))} \cdot (\mu - E(\mu(\underline{x}))^2)\}$$

La posterior naturalmente è una normale dato che la prior è una normale.

$$\pi(\mu|\underline{x}) \propto exp\{-\frac{1}{2(\sigma^2+\sigma_0^2)}(\sigma_0^2 n\mu - 2\mu n\overline{x}\sigma_0^2 + \mu^2\sigma^2 - 2\mu_0\mu\sigma^2)\}$$

Voglio isolare μ^2 per poi gestire il quadrato di un binomio in modo più intuitivo:

$$\propto exp\{-\frac{1}{2(\sigma^2+\sigma_0^2)}(n\sigma_0^2+\sigma^2)[\mu^2 - 2\mu\cdot\frac{n\overline{x}\sigma_0^2+\mu_0\sigma^2}{n\sigma_0^2+\sigma^2}]\}$$

$$\propto exp\{-\frac{1}{2}\frac{(n\sigma_0^2+\sigma^2)}{(\sigma^2+\sigma_0^2)}[\mu^2 - 2\mu\cdot\frac{n\overline{x}\sigma_0^2+\mu_0\sigma^2}{n\sigma_0^2+\sigma^2}]\}$$

Osserviamo che $Var(\mu|\underline{x}) = \frac{(\sigma^2+\sigma_0^2)}{(n\sigma_0^2+\sigma^2)}$ corrisponde alla varianza e $E(\mu|\underline{x}) = \frac{n\overline{x}\sigma_0^2+\mu_0\sigma^2}{n\sigma_0^2+\sigma^2}$ corrisponde alla media.

Il quadrato del binomio non ci interessa, poichè è una costante che non dipende da μ. Se lo voglio, aggiungo e tolgo all'esponente il seguente numero: $\pm(\mu|\underline{x}) = (\frac{n\overline{x}\sigma_0^2+\mu_0\sigma^2}{n\sigma_0^2+\sigma^2})^2$ in modo da ottenere:

$$\propto exp\{-\frac{1}{2}\frac{(n\sigma_0^2+\sigma^2)}{(\sigma^2+\sigma_0^2)}[\mu - \frac{n\overline{x}\sigma_0^2+\mu_0\sigma^2}{n\sigma_0^2+\sigma^2}]^2\}$$

La posterior ha quindi distribuzione:

$$\pi(\mu|\underline{x}) \sim N\Big(\frac{n\overline{x}\sigma_0^2+\mu_0\sigma^2}{n\sigma_0^2+\sigma^2}, \frac{(\sigma^2+\sigma_0^2)}{(n\sigma_0^2+\sigma^2)}\Big)$$

Osservando la media della posterior, vedo che c'è correlazione tra $\overline{x}$ e μ_0, cioè all'aumentare di n aumenta il peso della stima di massima verosimiglianza mentre all'aumentare di σ aumenta il peso della prior. Le varianze sono incrociate se cresce σ^2 mi fido di più della prior, se cresce σ_0^2 do più peso a $\overline{x}$.

Dire che la prior è molto variabile, vuol dire che è poco affidabile e quindi mi fido di più della stima di massima verosimiglianza.

3 Inferenza Bayesiana

Inferenza classica:

• Problemi strutturali: inferenza su θ
- stima puntuale
- stima intervallare (per regioni)
- verifica d'ipotesi

• Problemi previsivi: da un esperimento all'altro, se prendo un campione con caratteristica aleatoria, qual'è il risultato sensato?

$$\begin{cases} \underline{x} \Rightarrow \underline{x'} \\ n \Rightarrow n' \end{cases}$$

- stima puntuale
- stima intervallare (per regioni)
- verifica d'ipotesi

• Scelta e controllo del modello.

Soluzione Bayesiana ai problemi strutturali:

- stima puntuale: quale può essere un valore rappresentativo per θ. Qual'è la miglior approssimazione per stimare con un nucleo?
$\rightarrow E(\theta|\underline{x})$
$\rightarrow$ mediana a posteriori $(\theta|\underline{x})$
$\rightarrow$ moda $(\theta|\underline{x})$
$\rightarrow$ qualunque indice di sintesi di una distribuzione $(\theta|\underline{x})$ (vedremo più avanti che non è proprio vero)

- stima intervallare (per regioni): calcolo HPD e CS
$\rightarrow$ HPD (highest posterior density): fisso un numero positivo $h > 0$

$$S_h : \{\theta \in \Theta : \pi(\theta|\underline{x}) \geq h\}$$

con $h > h'$ ho che $S_h \subset S_{h'}$, cioè più abbasso il livello h e più mi accontento di più punti. Scelgo quindi un HPD tale che:

$$S_h : P(\theta \in S_h|\underline{x}) = 1 - \alpha$$

$\rightarrow$ CS (credible set): al fine di avere una coppia di affidabilità. Il credible set è un intervallo nel dominio di una distribuzione di probabilità a posteriori

utilizzato per avere stime intervallari. Credible set è l'analogo degli intervalli di confidenza della statistica frequentista.

3.1 Metodo Montecarlo

ESEMPIO: confronto tra proporzioni

Sono un consulente che deve dire se è meglio fare pubblicità per e-mail oppure per volantinaggio. Considero quindi n_1 unità statistiche estratte in modo casuale per e-mail e n_2 per volantini. I due campioni sono indipendenti.

n_1 (mandare una e-mail) avrà come esito della prova considerata x_1 a fronte della prima tecnica utilizzata. Avrò quindi n_1 prove indipendenti in $X_1 \sim Bernoulli(\theta)$, con θ ignoto. Avrò poi n_2 prove indipendenti in $X_2 \sim Bernoulli(\theta)$ sempre con θ ignoto.

$$H_0 : \quad \theta_1 = \theta_2$$

$ODDS_1 \rightarrow \frac{\theta_1}{1-\theta_1}$ a favore delle e-mail.
$ODDS_2 \rightarrow \frac{\theta_2}{1-\theta_2}$ a favore dei volantini.

ODDS: quanto sono disposto a scommettere sulle 2 tecniche?

$$OR\ RATIO = \frac{ODDS_1}{ODDS_2}$$

Se OR RATIO è uguale a 1, mi dice che per me è indifferente scegliere un metodo piuttosto che un altro. Il valore che assume è compreso tra 0 e ∞.

Per renderlo simmetrico ed interpretabile, utilizzo il logaritmo.

$$log(OR\ RATIO) = log\frac{ODDS_1}{ODDS_2}$$

- OR RATIO$<$ 0 ho asimmetria negativa.
- OR RATIO$=$ 0 sono indifferente.
- OR RATIO$>$ 0 ho asimmetria positiva.

Allora posso cambiare l'ipotesi $H_0 : OR = 0$ o $H_0 : \gamma = 0$
La posterior non si può calcolare analiticamente, quindi si deve utilizzare la **tecnica Montecarlo**.

$$\pi(\theta_1, \theta_2 | \underline{x_1}, \underline{x_2})$$

Ciò che voglio conoscere è la frequenza relativa dei soggetti che hanno comprato.
Sapendo che $\underline{x_1} = (1, 0, 0, 1, ..., 1)$ e $\underline{x_2} = (0, 0, 1, 1, ..., 0)$ mi basterà conoscere la statistica sufficiente $S_1 = \frac{\sum_{i=1}^{n_1} x_{i,1}}{n_1}$ e $S_2 = \frac{\sum_{i=1}^{n_2} x_{i,2}}{n_2}$.

Allora:

$$\pi(\theta_1, \theta_2|\underline{x_1}, \underline{x_2}) \propto \pi(\theta_1, \theta_2) \cdot f(\underline{x_1}, \underline{x_2}|\theta_1, \theta_2)$$

A priori so che θ_1 è indipendente da θ_2, quindi posso fattorizzare in prodotto di marginali tale che:

$$\pi(\theta_1, \theta_2) = \pi(\theta_1)\pi(\theta_2)$$

Con $\theta_1 = Beta(\alpha_1, \beta_1)$ e $\theta_2 = Beta(\alpha_2, \beta_2)$.

Dato che i campioni sono stati estratti in modo indipendente, allora

$$f(\underline{x_1}, \underline{x_2}|\theta_1, \theta_2) = f(\underline{x_1}|\theta_1) \cdot f(\underline{x_2}|\theta_2)$$

Ho il prodotto di 2 verosimiglianze fino a trovare $\rightarrow$ $\theta_1|\underline{x_1}$ indipendente da $\theta_2|\underline{x_2}$, come se facessi 2 modelli $Beta$ separati a posteriori.

$$\theta_1|\underline{x_1} \sim Beta(\alpha_1 + S_1, \beta_1 + n_1 - S_1)$$

$$\theta_2|\underline{x_2} \sim Beta(\alpha_2 + S_2, \beta_2 + n_2 - S_2)$$

Posso ora fare inferenza su θ senza problemi. Dato che la legge di distribuzione di γ non è calcolabile analiticamente, uso il metodo Montecarlo.
Mi serve quindi un campione sufficientemente grande di determinazioni indipendenti e identicamente distribuiti della posterior.

$$\gamma_1, \gamma_2, ..., \gamma_m$$

Con m che corrisponde ad un numero intero e se $m \rightarrow \infty$ so che $\frac{\sum_{i=1}^{m} \gamma_i}{n} \rightarrow E(\gamma|\underline{x_1}, \underline{x_2})$ è una stima consistente per la media della posterior (con $\pi(\gamma|\underline{x_1}, \underline{x_2})$ ignota).

Come costruisco il campione Montecarlo per γ?
Sapendo che γ è nota ed è uguale a:

$$\gamma = log\left(\frac{\frac{\theta_1}{1-\theta_1}}{\frac{\theta_2}{1-\theta_2}}\right)$$

Avendo ora $\pi(\theta_1|\underline{x_1})$ nota e $\pi(\theta_2|\underline{x_2})$ nota, posso estrarre campioni casuali, ottenendo:

$$\theta_{1,1}, ..., \theta_{1,m}$$

$$\theta_{2,1}, ..., \theta_{2,m}$$

I campioni casuali sono indipendenti e identicamente distribuiti da $\pi(\theta_1|\underline{x_1})$ per la prima e $\pi(\theta_2|\underline{x_2})$ per la seconda.

Quindi il campione Montecarlo diventa:

$$\gamma_i = log\left(\frac{\frac{\theta_{1,i}}{1-\theta_{1,i}}}{\frac{\theta_{2,i}}{1-\theta_{2,i}}}\right)$$

3.2 Inferenza previsiva

A partire dalla posterior si può fare inferenza strutturale (cioè sui parametri). Vediamo ora come fare inferenza previsiva.

Eravamo partiti dal modello statistico di base $(x, f(x,\theta), \underline{\theta} \in \Theta)$, facendo n prove.
Se facessi un secondo esperimento o modello statistico ottengo $(x, f'(x',\theta), \underline{\theta} \in \Theta)$, facendo n' prove. Il parametro θ rimane lo stesso, DEVE essere lo stesso. Se non posso ritenere che sia lo stesso, allora non posso fare la previsione.

Ho la prima $n-upla$ di valori noti e voglio fare previsione sulla $n-upla$ di $\underline{x'}$ ignota.

Fino ad ora avevamo:

$$\underline{x} \Rightarrow f(\underline{x}|\theta)$$

$$\theta \Rightarrow \pi(\theta)$$

Osserviamo che ci sono solo 2 oggetti: $\Psi(\underline{x}, \theta)$

Ora abbiamo invece 3 oggetti, poichè si aggiunge:

$$\underline{x'} \Rightarrow f'(\underline{x'}|\theta)$$

ottenendo così: $\Psi(\underline{x}, \underline{x'}, \theta)$

Avendo ora a disposizione i 3 elementi: $f(\underline{x}|\theta)$, $\pi(\theta)$ e $f'(\underline{x'}|\theta)$ posso farne il prodotto e ottenere

$$f(\underline{x}|\theta) \cdot f'(\underline{x'}|\theta) \cdot \pi(\theta)$$

da cui ricavo

$$g(\underline{x}, \underline{x'}|\theta) \cdot \pi(\theta)$$

che corrisponde a

$$\Psi(\underline{x}, \underline{x}', \theta)$$

Questo si può fare se c'è indipendenza condizionatamente a θ, allora mi basta fare 2 estrazioni indipendenti dei campioni e θ rimane lo stesso.

Calcolare la legge congiunta sotto certe ipotesi è possibile.
Abbiamo θ ignoto, $\underline{x}$ noto (corrisponde all'esito del primo esperimento) e $\underline{x'} = S'$ (non so quanto è, ma è costante).
Dato che θ è ignoto, non avrebbe senso inserirlo nella previsione, a me interessa $m(\underline{x}'|\underline{x})$
Precedentemente avevamo visto $m(\underline{x})$ che corrispondeva alla **predittiva iniziale**, cioè la marginale di $\underline{x}$
$m(\underline{x}'|\underline{x})$ invece è la **predittiva finale** (posterior predicted).

Per calcolare $m(\underline{x}^{'}|\underline{x})$ posso partire dalla distribuzione congiunta condizionata $h(\underline{x}^{'}, \theta|\underline{x})$, vediamo come:

$$m(\underline{x}^{'}|\underline{x}) = \int_{\Theta} h(\underline{x}^{'}, \theta|\underline{x})\, d\theta = \int_{\Theta} f^{'}(\underline{x}^{'}|\theta)\pi(\theta|\underline{x})\, d\theta$$

viene sfruttato il fatto che i due esperimenti, condizionatamente a θ, sono indipendenti.
A livello di calcoli c'è una via più veloce:

$$m(\underline{x}^{'}|\underline{x}) = \frac{\Psi(\underline{x}^{'}, \underline{x})}{m(\underline{x})} = \frac{\int_{\Theta} \Psi(\underline{x}', \underline{x}, \theta)\, d\theta}{\int_{\Theta} \Psi(\underline{x}, \theta)} = \frac{\int_{\Theta} f(\underline{x}|\theta) f^{'}(\underline{x}^{'}|\theta)\pi(\theta)\, d\theta}{\int_{\Theta} f(\underline{x}|\theta)\pi(\theta)\, d\theta}$$

Esercizio Binomiale-Beta
Vengono fatte $n = 10$ prove indipendenti su un fenomeno dicotomico e il numero di imprese con soglia di titoli rischiosi elevata é $s = 2$. Se faccio altre $n^{'} = 20$ prove e assunto che θ sia lo stesso, quale sarebbe una previsione per $s^{'}$?
In ambito classico calcolerei $E(s^{'}) = n^{'}\theta = n^{'}\hat{\theta} = 4$.
In ambito bayesiano mi serve $f(\underline{x}|\theta) = \theta^{s}(1-\theta)^{n-s}$, dove s é il numero di successi nel primo campione. Per il secondo esperimento bisogna calcolare $f^{'}(s^{'}|\theta) = \binom{n^{'}}{s^{'}}\theta^{s^{'}}(1-\theta)^{n^{'}-s^{'}}$ ma $s^{'}$ e $n^{'}$ sono ignoti.
Come prior prendo una variabile casuale $Uniforme \rightarrow \pi(\theta) = 1$.
Allora

$$m(\underline{x}^{'}|\underline{x}) = \frac{\int_{\Theta} f^{'}(s^{'}|\theta) f(\underline{x}|\theta)\pi(\theta)\, d\theta}{\int_{\Theta} f(\underline{x}|\theta)\pi(\theta)\, d\theta}$$

$$= \binom{n^{'}}{s^{'}} \frac{B(s + s^{'} + 1, n + n^{'} - s - s^{'} + 1)}{B(s+1, n-s+1)}$$

$$= \binom{20}{s'} \frac{B(s' + 3, 29 - s')}{B(3, 9)}$$

questa è la funzione di probabilità di una $BETA\ BINOMIALE$ con supporto $s' = 0, 1, ..., n'$.
Per il momento come previsione bayesiana per s' va bene qualsiasi sintesi, più avanti vedremo un approccio decisionale.

Valore atteso:
Attraverso la formula già calcolata:

$$E(s'|\underline{x}) = n' \cdot \frac{s+1}{n+2} = (\sum_{s'=0}^{n'} s' \psi(s')) = 20 \cdot \frac{1}{4} = 5$$

Previsione puntuale non coerente con quella intuitiva fatta in via classica.

4 Elicitazione della prior

Ci sono diverse tecniche per elicitare la prior:
- assegnazione diretta
- distribuzioni non informative
- distribuzioni coniugate al modello
- elicitazione tramite predittiva iniziale

4.1 Assegnazione diretta

Si utilizza quando le informazioni a priori che si hanno sul parametro sono molto abbondanti (si rischia di influenzare pesantemente la posterior). In pratica si assegna direttamente e soggettivamente le probabilità ai valori di θ (se è una variabile casuale discreta): quindi mi vengono date **k** prior differenti (k indica la cardinalità del supporto del parametro).

$$\theta_1, \theta_2, ..., \theta_k \text{ (con k= } \#\Theta)$$

$$\pi(\theta_1), ..., \pi(\theta_k)$$

Qualcuno soggettivamente mi fornisce questi concetti.
Nel caso in cui il supporto non è finito si assegna soggettivamente una probabilità a priori a diversi intervalli del parametro θ.

4.2 Distribuzioni non informative

Non inseriscono soggettività ma discendono da meccanismi automatici, viene fornita una regola di elicitazione.

4.2.1 Laplace

Viene assegnata la stessa probabilità (o densità) a priori a ciascun valore di θ, quindi se ad esempio la cardinalità del supporto è k (finita), la prior per ogni valore del parametro è $\frac{1}{k}$. L'uniforme è un esempio di queste prior. Si incontrano problemi quando Θ non é limitato perchè non esiste una costante che se integrata sul supporto da come risultato 1; dunque non posso usare Laplace.

DEFINIZIONE:
se l'integrale di una priori sul suo supporto diverge, allora $\pi(\theta)$ non è una densità.

$$\int_{\Theta} \pi(\theta)\ d\theta = \infty$$

Si dice che la prior è **impropria**.
In ambito bayesiano si può usare comunque purchè la posterior sia propria.

ESEMPIO
Si fanno n prove indipendenti in X, dove

$$f(X|\theta) \propto \theta^4 x^3 e^{-\theta x} \quad \text{con } x > 0,\ \theta > 0$$

Domanda: ha senso scegliere una prior di Laplace?

Se integro su $\mathbb{R}^+$ una costante questa diverge, dunque non esiste alcuna C tale che il suo integrale valutato sul supporto converga. Allora $\pi_L(\theta)$ é impropria (per comodità prendo $C = 1$). quindi:

$$\nexists\ C: \text{ tale che } \pi(\theta) = C \text{ sia propria}$$

La posterior è propria?

$$\pi(\theta|\underline{x}) \propto \pi(\theta) f(\underline{x}|\theta) \propto 1 \cdot \theta^{4n} (\prod_{i=1}^{n} x_i)^3 \cdot e^{\theta \sum x_i}$$

Ora se riconosco il nucleo di una variabile casuale nota sono a posto perchè so come si distribuisce la posterior, altrimenti devo capire se è normalizzabile (e dunque propria) oppure se non lo è (e dunque impropria).

$$\begin{cases} \text{se normalizzabile} \Rightarrow \text{propria} \\ \text{se non normalizzabile} \Rightarrow \text{impropria} \end{cases}$$

In questo caso ho il nucleo di una $Gamma(4n + 1, \sum x_i)$.

Risposta: Si, ha senso utilizzare la tecnica di Laplace perchè nonostante la prior sia impropria, la posterior è propria.

DEFINIZIONE:
data la variabile X e $f(x;\theta)$, con n prove indipendenti ottengo $f(\underline{x}|\theta)$, ma potrei essere interessato a fare inferenza su $\lambda = g(\theta)$. Se la funzione è biunivoca (è una funzione regolare che posso invertire) allora ho una regola di elicitazione (di coerenza) con la quale posso ottenere $\pi(\theta)$, ma come deduco la prior di λ?
Una prima via consiste nell'elicitare e poi trasformare:

$$\pi^*(\lambda) = \pi(g^{-1}(\lambda)) \left| \frac{\delta g^{-1}(\lambda)}{\delta \lambda} \right|$$

In alternativa potrei prima riparametrizzare e poi elicitare: parto da $(\mathbf{X}, f(x|\theta), \theta \in \Theta)$ e arrivo a $(\mathbf{X}, f(x|\lambda), \lambda \in \Lambda)$, da cui calcolo $\tilde{\pi}(\lambda)$.
C'è **invarianza** se: $\pi^*(\lambda) = \tilde{\pi}(\lambda)$, cioè se è indifferente elicitare e poi trasformare o fare prima la trasformazione e poi l'elicitazione.

ESERCIZIO
Dato

$$f_x(x|\theta) = \frac{\theta^3 e^{-\frac{\theta}{x}}}{2x^4} \text{ con } x > 0,\ \theta > 0$$

Domanda: calcolare $\pi_L(\theta)$ (prior di Laplace) e decidere se è invariante rispetto alla riparametrizzazione $\lambda = \theta + 1$.

Usando la regola di elicitazione ottengo $\pi_L(\theta) = c = 1$ che dunque è **impropria**; ora applicando la trasformazione ricavo

$$\lambda = g(\theta) = \theta + 1 \rightarrow \theta = g^{-1}(\lambda) = \lambda - 1$$

Allora $\pi^*(\lambda) = \pi(g^{-1}(\lambda)) \left| \frac{\delta g^{-1}(\lambda)}{\delta \lambda} \right| = 1 \cdot 1 = 1$, cioè con il primo metodo la prior è pari a 1.
Ora parto dalla riparametrizzazione:

$$f(x|g^{-1}(\lambda)) = \frac{(\lambda-1)^3 e^{-\frac{(\lambda-1)}{x}}}{2x^4}$$

ed elicito $\rightarrow \tilde{\pi}(\lambda) = 1$.
Risposta: ho ottenuto $\pi^*(\lambda) = \tilde{\pi}(\lambda) \rightarrow$ c'é invarianza. (semplicemente so che se la funzione di θ è una traslazione, allora c'è invarianza)

ESERCIZIO
Sia $X \sim Po(\theta)$ e $\lambda = g(\theta) = \frac{1}{\theta}$
Domanda: utilizzando la regola di Laplace c'è invarianza?

Prima di tutto calcolo $f(\underline{x}|\theta) = \frac{e^{-\theta}\theta^x}{x!}$
Se elicito (noto subito che la prior è impropria) e poi trasformo ottengo $\pi^*(\lambda) \propto \frac{1}{\lambda^2}$, mentre partendo dall'elicitazione e facendo successivamente la trasformazione ottengo $\tilde{\pi}(\lambda) = c$

4.2.2 Jeffreys

Associa una regola di elicitazione non soggettiva

$$\pi_J(\theta) \propto \sqrt{I_A(\theta)}$$

dove $I_A(\theta)$ è l'informazione attesa di Fisher.

$$I_A(\theta) = E\left[\frac{\delta log(f(X;\theta))}{\delta\theta}\right]^2 = -E\left[\frac{\delta^2 log(f(X;\theta))}{\delta\theta^2}\right]$$

Dunque una volta che si ha una prova se ne calcola l'informazione attesa di Fisher e quanto ottenuto è il nucleo della prior.
Anche $\pi_J(\theta)$ è automatica come Laplace, può essere propria o impropria e inoltre è **SEMPRE invariante**.

ESERCIZIO 1
Sia $X \sim Po(\theta)$ e $n = 1$.
Domanda: calcolare $\pi_J(\theta)$.

$$f(x;\theta) = \frac{e^{-\theta}\theta^x}{x!}$$

$$\Rightarrow I_A(\theta) = -E[\frac{\delta^2(-\theta + xlog(\theta) - log(x!))}{\delta\theta^2}] = E[\frac{x}{\theta^2}] = \frac{1}{\theta^2}\theta = \frac{1}{\theta}$$

Risposta: $\pi_J(\theta) \propto \sqrt{\frac{1}{\theta}} = \theta^{-\frac{1}{2}}$.
Dato che c'è invarianza posso dedurre la prior per trasformazioni, è sufficiente calcolare $\pi^*(\lambda)$.
Domanda: E' propria?
La prior è **sempre** invariante e in questo caso è impropria. La trasformata $\lambda = \frac{1}{\theta} = g(\theta)$ è trasformata biunivoca.
Ora vediamo le due possibili vie:

- $\theta = \frac{1}{\lambda} \rightarrow \delta\theta = |-\frac{1}{\lambda^2}|$. Allora la prior che ottengo è

$$\pi_J^*(\lambda) \propto (\frac{1}{\lambda})^{-\frac{1}{2}}\frac{1}{\lambda^2} = \lambda^{-\frac{3}{2}}$$

- parto dalla riparametrizzazione ho $f(x|\theta) = \frac{e^{-\theta}\theta^x}{x!} \rightarrow f(x|\lambda) = \frac{e^{-\frac{1}{\lambda}}\frac{1}{\lambda}^x}{x!}$ con $x = 0, 1, ..$ e $\theta > 0 \rightarrow \lambda > 0$.

Devo calcolare $\tilde{\pi_J}(\lambda) \propto \sqrt{I_A(\lambda)}$, allora

$$log(f(x|\lambda)) = -\frac{1}{\lambda} - xlog(\lambda) - log(x!) \rightarrow \frac{\delta^2 log(f(x|\lambda))}{\delta\lambda^2} = -\frac{2}{\lambda^3} + \frac{x}{\lambda^2}$$

Ora cambio il segno dell'equazione e calcolo il valore atteso (rispetto a x)

$$I_A(\lambda) = \frac{2}{\lambda^3} - \frac{1}{\lambda^2}E(X) = \frac{1}{\lambda^3}$$

Allora $\tilde{\pi_J}(\lambda) \propto \sqrt{\frac{1}{\lambda^3}} = \lambda^{-\frac{3}{2}} = \pi_J^*(\lambda)$

Con questa verifica ho dimostrato che la prior di Jeffreys è invariante.

ESERCIZIO 2

Sia $X \sim Ber(\theta)$ e $n = 1$; $\pi_J(\theta) \propto \sqrt{I_A(\theta)}$ è normalizzabile.

Domanda: perchè mi basta prendere l'informazione di una prova?
Risposta: si ottiene lo stesso risultato se prendo un campione più grosso perchè facendo n prove indipendenti ho n volte l'informazione di una prova. Ma poichè n è una costante e Jeffreys mi da solo il nucleo utilizzando una proporzione, la prior che si ricava è la stessa (e comunque non avrebbe senso avere una prior che dipende dall'ampiezza campionaria).

Domanda: cosa succede se θ non é uno scalare?
Risposta: se ho il vettore $\underline{\theta} = (\theta_1, ..., \theta_k)$ l'informazione di Fisher diventa una matrice simmetrica $k \cdot k$ in cui i singoli elementi sono

$$I_{i,j} = -E\left[\frac{\delta^2 log(f(x|\underline{\theta}))}{\delta\theta_i\delta\theta_j}\right]$$

e quindi

$$\pi_J(\underline{\theta}) \propto \sqrt{det(I_A(\underline{\theta}))}$$

Soluzione:

$\pi_J(\theta) \propto \frac{1}{\sqrt{\theta(1-\theta)}}$ è normalizzabile $\pi_J(\theta) = \frac{1}{\pi\sqrt{\theta(1-\theta)}}$ è propria.

Osservazione: $\pi_J(\theta) \propto \frac{1}{\pi}\theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} \rightarrow$ riconosco il nucleo di una $Beta(\frac{1}{2}, \frac{1}{2})$.

Inoltre più sono piccoli i parametri α e β meno importanza viene data alla prior (e conta di più il campione); si vede chiaramente che il meccanismo di

Jeffreys è davvero non informativo perchè mette poca rilevanza alla prior.

Esercizio (multiparametrico)
Data una $X \sim N(\mu, \sigma)$ con parametri ignoti.
N.B. Posso usare indifferentemente σ o σ^2 **se voglio una a priori non informativa di Jeffreys** perchè c'è invarianza e la trasformazione è biunivoca.

$$\pi_J(\mu, \sigma) \propto \sqrt{det(I_A(\mu, \sigma))}$$

Vediamo la condizionata

$$f(X|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \rightarrow log(f(X|\mu, \sigma))$$

Ora devo calcolare le derivate seconde e miste:

$$\frac{\delta^2 log(f)}{\delta \mu^2} = ... = -\frac{1}{\sigma^2}$$

$$\frac{\delta^2 log(f)}{\delta \mu \delta \sigma} = ... = -\frac{2(x-\mu)}{\sigma^3}$$

$$\frac{\delta^2 log(f)}{\delta \sigma^2} = ... = \frac{1}{\sigma^2} - \frac{3(x-\mu)^2}{\sigma^4}$$

$$\rightarrow I_A(\mu, \sigma) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}$$

quindi

$$\pi_J(\mu, \sigma) \propto \sqrt{det(I_A(\mu, \sigma))} \propto \frac{1}{\sigma^2}$$

è la distribuzione a priori invariante non informativa di Jeffreys per (μ, σ).
E' propria (se si la normalizzo, altrimenti no)?

$$\int_R \int_{R^+} \frac{1}{\sigma^2} \, d\mu \, d\sigma \rightarrow \text{ non converge}$$

la prior non è propria.

Osservazione: l'elicitazione della prior di Jeffreys nel caso multiparametrico è problematica, un'ipotesi spesso utilizzata è l'indipendenza a priori

$$\pi(\underline{\theta}) = \pi_1(\theta_1) \cdot ... \cdot \pi_k(\theta_k) \propto \sqrt{det(I_A(\underline{\theta}))} \neq \sqrt{det(I_A(\theta_1))} \cdot ... \cdot \sqrt{det(I_A(\theta_k))}$$

in generale non coincidono.

4.2.3 Vague

Si utilizzano prior di una famiglia di distribuzioni imponendo che la varianza sia molto elevata.

ESEMPIO
Con $\theta \in \mathbb{R}$ prendo $\theta \sim N(\mu, 10^6)$. In questo caso la forma della distribuzione è molto simile ad un'uniforme, quindi presi intervalli centrali le probabilitá sono pressochè uguali e la prior è propria.

4.3 Prior coniugate

Necessitano di una classe parametrica di distribuzioni $\mathcal{D} = \{\pi_\alpha(\theta), \alpha \in A\}$ ed hanno tutte la stessa forma funzionale.

DEFINIZIONE: dato un modello $(X, f(x;\theta), \theta \in \Theta)$ allora la classe parametrica $\mathcal{D}$ di distribuzioni per θ si dice **coniugata** al modello se scelta in $\mathcal{D}$ la priori anche la posteriori vi appartiene per ogni x.

Questo facilita gli aspetti computazionali perchè se so come è coniugata la priori conosco già anche la distribuzioni della posteriori; inoltre anche l'aggiornamento dell'iperparametro della prior e della posterior è facile $\alpha \to \alpha_p$.

Modello base	**Classe coniugata**	**Aggiornamento iper-parametri**
$Be(\theta)$	$Beta(\alpha, \beta)$	$Beta(\alpha + \text{s}, \beta + \text{n+s})$
$Po(\theta)$	$Gamma(\alpha, \beta)$	$Gamma(\alpha + \text{s}, \beta + \text{n})$
$ExpNeg(\theta)$	$Gamma(\alpha, \beta)$	$Gamma(\alpha + \text{n}, \beta + \text{s})$
$N(\mu, \sigma^2 = \text{nota})$	$N(\mu_0, \sigma_0^2)$	$N(\frac{\mu_0\sigma^2 + n\bar{x}\sigma_0^2}{\sigma_0^2 n + \sigma^2}, \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2})$
$N(\mu = \text{nota}, \sigma^2)$	$GaInv(\alpha, \beta)$	$GaInv(\alpha + \frac{n}{2}, \beta + \frac{(\sum x_i - \mu)^2}{2})$
$N(\mu = \text{nota}, \frac{1}{\sigma^2})$	$Gamma(\alpha, \beta)$	$Ga(\alpha + \frac{n}{2}, \beta + \frac{(\sum x_i - \mu)^2}{2})$
$Uniforme(0, \theta)$	$Pareto(\alpha, \beta)$	$Pareto(\alpha + \text{n}, Max\{\beta, X_{(n)}\})$

La variabile casuale Gamma Inversa

$$X \sim GaInv(\alpha, \beta)$$

$$f(x;\alpha,\beta)=\frac{\beta^\alpha}{\Gamma(\alpha)}\frac{1}{x^{\alpha+1}}e^{-\frac{\beta}{x}} \text{ con } \alpha,\beta>0 \text{ e } x>0$$

Allora:

$$Y=\frac{1}{X}\sim Gamma(\alpha,\beta)$$

La variabile casuale di Pareto

$$X\sim Pareto(\alpha,\beta)$$
$$f(x;\alpha,\beta)=\alpha\beta^\alpha\frac{1}{x^{\alpha+1}}I_{[\beta,\infty](x)} \text{ con } \alpha,\beta>0$$

ESERCIZIO
La famiglia parametrica di Pareto è coniugata al modello uniforme.

$$X\sim U(0,\theta)$$

faccio n prove indipendenti in x, $f(x|\theta)=\frac{1}{\theta}$ con $0<x<\theta$.

$$f(\underline{x}|\theta)=\prod_{i=1}^{n}f(x_i|\theta)=\frac{1}{\theta^n} \text{ con } x_{(n)}\leq\theta\rightarrow x_i\leq\theta\ \forall i$$

Per calcolare la posterior mi interessa capire dove si muove θ, cioè $\theta\geq x_{(n)}$.

$$L(\theta)=\frac{1}{\theta^n}I_{[x_{(n)},\infty]}(\theta)\rightarrow\pi_{\alpha,\beta}(\theta)=\alpha\beta^\alpha\frac{1}{\theta^{\alpha+1}}I_{[max(x_{(n)},\beta),\infty]}(\theta)$$

Allora

$$\pi(\theta|\underline{x})\propto\frac{1}{\theta^n}I_{[]x_{(n)},\infty]}(\theta)\frac{1}{\theta^{\alpha+1}}I_{[\beta,\infty]}(\theta)=\frac{1}{\theta^{n+\alpha+1}}I_{[max(x_{(n)},\beta),\infty]}(\theta)$$

questo è il modello di una

$$Pareto(\alpha+n,max\{\beta,x_{(n)}\})$$

ESEMPIO:
Dato un modello $X\sim N(\mu,\sigma^2)$ con μ nota (corrisponde quindi ad un numero che può essere eliminato nel calcolo della posterior) e σ^2 non nota, avrà come classe coniugata una *Gamma Inversa*.

Consideriamo quindi la **funzione di verosimiglianza** di una normale con media nota e varianza ignota.

$$f(\underline{x}|\sigma^2)=(\sigma^2)^{-\frac{n}{2}}\cdot e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

Ho deciso di trascurare $\frac{1}{\sqrt{2\pi}}$ poichè è una costante che non dipende dal parametro ignoto.

La **prior** da una *Gamma Inversa* sarà quindi:

$$\frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{1}{(\sigma^2)^{\alpha+1}} \cdot e^{-\frac{\beta}{\sigma^2}}$$

Nella prior trascuro la parte iniziale $\frac{\beta^\alpha}{\Gamma(\alpha)}$ poichè non dipende dal parametro σ^2.
Noto che la funzione di verosimiglianza di una *Normale* ha una struttura molto simile alla prior di una *Gamma Inversa*, perciò deduco che la posterior sarà una *Gamma Inversa*.

Posterior:

$$\pi(\sigma^2|\underline{x}) = (\sigma^2)^{-\frac{n}{2}-\alpha-1} \cdot e^{-\frac{1}{\sigma^2}[+\frac{\sum_{i=1}^n (x_i-\mu)^2}{2}+\beta]}$$

Nella posterior ho un nucleo noto $\sim GaInv(\alpha + \frac{n}{2}, \beta + \frac{\sum_{i=1}^n (x_i-\mu)^2}{2})$.

Risultato utile e costruttivo:
Dal modello voglio trovare la famiglia coniugata. In generale non conosco la coniugata, perciò voglio capirla senza che nessuno me la dica.

Il risultato è sulla *Famiglia Esponenziale* :

$X \in$ *Famiglia Esponenziale* se

$$f(x;\theta) = e^{\theta x - C(\theta)} \cdot D(x)$$

Ho quindi un solo parametro canonico, naturale (θ), due funzioni C e D e x osservazione naturale.
Il supporto di x non dipende da θ.
Se il modello appartiene alla *famiglia esponenziale* allora:

Teorema:
Sia $x \in$ *Famiglia Esponenziale*, allora la prior è:

$$\pi(\theta) \propto e^{\eta_1\theta - \eta_2 C(\theta)}$$

(η_1 ed η_2 sono iperparametri che non si muovono ovunque e garantiscono che $\pi(\theta)$ sia propria, quindi normalizzabile).

La prior $\pi(\theta)$ è coniugata al modello.

Dimostrazione:
se $\pi(\theta) \propto e^{\eta_1\theta-\eta_2 C(\theta)}$ allora la posterior appartiene alla stessa famiglia:

$$\pi(\theta|\underline{x}) \propto e^{\theta\sum_{i=1}^n x_i - nC(\theta)} \cdot \prod_{i=1}^n D(x_i) \cdot e^{\eta_1\theta-\eta_2 C(\theta)} \propto e^{\theta(\sum_{i=1}^n x_i-\eta_1)-C(\theta)[n-\eta_2]}$$

$$\pi(\theta|\underline{x}) \sim (\eta_1 + \sum_{i=1}^n x_i, \eta_2 + n)$$

$$\pi(\theta) \sim (\eta_1, \eta_2)$$

Osservazione:
Se $f(x;\theta) = e^{A(\theta)\cdot B(x)-C(\theta)} \cdot D(x)$ allora il risultato va bene e cercherò di ricondurmi sempre a questa forma.
In questo caso non avrò $\sum x_i$ ma la trasformata $\sum B(x_i)$.
Se il parametro θ non è canonico, naturale, cioè che compare una funzione di θ, allora il parametro canonico sarà $A(\theta) = \lambda$ che può essere una qualsiasi funzione (ad esempio $log(\theta)$).
Il risultato continua a valere per λ.

Trovo la prior coniugata per $A(\theta) = \lambda$ (esempio $log(\theta) = \lambda$):
- riparametrizzo in λ, cioè sostituisco la funzione in λ.
- applico il teorema (che mi dice come trovare la coniugata per λ).
- se invertibile, ne deduco la priori per θ e se non invertibile non faccio niente.

Esercizio su Bernoulli:
$X \sim Bernoulli(\theta)$ allora voglio sapere cosa sarà $\pi(\theta)$ coniugata (avendola già studiata, sappiamo a priori che sarà una $Beta$, ma facciamo come se non sapessimo nulla).

$X \in Famiglia\ Esponenziale$?

$$f(x;\theta) = \theta^x(1-\theta)^{1-x} \quad x = 0, 1 \ e \ 0 < \theta < 1$$

che posso riscrivere in forma più intuitiva come:

$$f(x;\theta) = e^{xlog(\theta)+(1-x)log(1-\theta)}$$

in questa forma non è ancora esponenziale, quindi:

$$e^{x[log(\theta)-log(1-\theta)]+log(1-\theta)}$$

- $log(\theta) - log(1-\theta)$ corrisponde all'ODDS $= log\frac{\theta}{1-\theta}$
- $x[log\frac{\theta}{1-\theta}] = x \cdot A(\theta) = x \cdot \lambda$ *con* λ *parametro canonico*
- $-log(1-\theta) = C(\theta)$

Devo ora riparametrizzare:
- Dato che $\lambda = log\frac{\theta}{1-\theta}$, riparametrizzando ho $\theta = \frac{e^\lambda}{1-e^\lambda}$
- $1-\theta = \frac{1}{1+e^\lambda}$
- $log(1-\theta) = -(C(\theta)) = -log(1+e^\lambda)$

La funzione di verosimiglianza è:

$$f(x;\theta) = e^{x\cdot\lambda - log(1+e^\lambda)}$$

Applico il teorema: le prior per λ coniugate al modello:

$$\pi(\lambda) \propto e^{\eta_1\lambda - \eta_2 log(1+e^\lambda)}$$

Voglio conoscere la prior di θ, quindi trovo il differenziale:

$$\lambda = log\frac{\theta}{1-\theta} = log(\theta) - log(1-\theta)$$

$$|\delta\lambda| = \left|\frac{1}{\theta} + \frac{1}{1-\theta}\right| d\theta = \frac{1}{\theta(1-\theta)}\ d\theta$$

La prior di θ sarà quindi:

$$\pi(\theta) \propto e^{\eta_1\cdot(log\frac{\theta}{1-\theta}) - \eta_2\cdot log(1-\theta)} \cdot \frac{1}{\theta(1-\theta)}$$

La prior coniugata di λ è $e^{\eta_1\cdot(log\frac{\theta}{1-\theta}) - \eta_2\cdot log(1-\theta)} = \pi(A(\theta))$

$$(\frac{\theta}{1-\theta})^{\eta_1} \cdot (1-\theta)^{\eta_2} \cdot \frac{1}{\theta(1-\theta)} = (\theta)^{\eta_1 - 1} \cdot (1-\theta)^{\eta_2 - \eta_1 - 1}$$

che corrisponde al nucleo di una $Beta(\eta_1, \eta_2 - \eta_1)$ con gli iperparametri $\eta_1 > 0$ e $\eta_2 - \eta_1 > 0$

ESERCIZIO SU POISSON

Esercizio su normale:
$X \sim N(\mu, \sigma^2)$ (sappiamo già che θ è una $Normale$).
La funzione di verosimiglianza è:

$$f(x;\mu) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(x-\mu)^2} \propto e^{x\cdot\mu - \frac{\mu^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}$$

con $x \cdot \mu - \frac{\mu^2}{2}$ che corrisponde a $x \cdot \mu - C(\mu)$ e $\frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} = D(x)$
μ è il parametro canonico e $C(\mu) = \frac{\mu^2}{2}$.

$$\pi(\mu) \; coniugata \; \propto e^{\eta_1 \mu - \eta_2 \frac{\mu^2}{2}}$$

che corrisponde al nucleo di una normale $N \sim (\frac{\eta_1}{\eta_2}, \frac{1}{\eta_2})$
$\frac{\eta_1}{\eta_2}$ e $\frac{1}{\eta_2}$ sono numeri, valori noti.

4.4 Metodo di scelta degli iperparametri

Se $\pi \in \mathscr{D}$ abbiamo un problema sulla scelta degli iperparametri.

1) Diretta
Non ho nulla da dire perchè effettuo una scelta **diretta** (ad esempio un esperto mi da i momenti, i quantili da cui deduco gli iperparametri).

2) Assegnazione gerarchica
Scelgo la forma funzionale della prior:

$$\pi \in \mathscr{D} = \{\pi(\theta, \underline{\alpha}); \alpha \in A\}$$

con $\underline{\alpha}$ corrispondente ad uno scalare o ad un vettore di iperparametri.

Multifase: ipotizza che anche l'iperparametro o vettore di iperparametri $\underline{\alpha}$ sia una variabile casuale $\Rightarrow \pi(\underline{\alpha})$.
Tipicamente ci si ferma alla seconda fase, scegliendo per l'iperparametro $\underline{\alpha}$ **Hyper prior non informative**.

Cerco $\pi(\theta) \Leftarrow$ elicitazione in due fasi:
Prima fase: $\pi(\theta|\underline{\alpha})$ vettore condizionato ad un'altra variabile casuale.
Seconda fase: $\pi(\underline{\alpha})$.

$$\pi(\theta) = \int_A \pi(\theta|\underline{\alpha}) \cdot \pi(\underline{\alpha}) \; d\underline{\alpha}$$

Esercizio tema d'esame
$X \to \mathscr{X} = [0, \theta]$ (supporto)
$\underline{x} = (1.5, 2, 1.9, 2.3, 0.3, 2.5, 2.8)$ e $n = 7$
θ è almeno 2.8 che corrisponde al massimo, cioè $\theta \geq x_{(n)} = 2.8$

$X \sim Uniforme(0, \theta)$
$\pi(\theta) \propto \theta^{-4}$, $\theta > \beta \; con \; \beta = 2$

Domanda 1: calcolo la posterior (sarà una $Pareto(10, 2.8)$)
Domanda 3: supponendo $\pi(\theta) \propto \theta^{-\alpha-1}$ $, \theta > 1$ *e* $\alpha > 0$, scelta una hyper prior di Laplace per α, si imposti il calcolo della posterior $\pi(\theta|\underline{x})$.

$$\pi(\theta|\underline{x}) \propto \pi(\theta) f(\underline{x}|\theta)$$

La prior $\pi(\theta|\underline{x})$ è ignota, però sappiamo che sarà $\propto Pareto(\alpha, 2)$
$\pi(\theta|\alpha) = \alpha 2^{\alpha} \theta^{-\alpha-1} I_{[2,+\infty]}(\theta)$ dato che è nota.
$\pi(\underline{\alpha}) = 1$ (so che è impropria e quindi scelgo 1).
Ricavo quindi la prior di θ:

$$\pi(\theta) = \int_0^{\infty} \alpha 2^{\alpha} \theta^{-\alpha-1} I_{[2,+\infty]}(\theta) \cdot 1 \; d\underline{\alpha}$$

$$\frac{I_{[2,+\infty]}(\theta)}{\theta} \int \alpha (\frac{2}{\theta})^{\alpha} \; d\alpha$$

$$.....svolgo.....$$

$$\frac{1}{\theta} I_{[2,+\infty]}(\theta) \frac{1}{(ln\frac{2}{\theta})^2}$$

3) Assegnazione empirica
$\pi(\theta; \underline{\alpha})$ ad $\underline{\alpha}$ assegno un valore o un vettore di valori multidimensionale, calcolato a partire dal campione.

$$h(\underline{x}) \Leftarrow \underline{\alpha^0}$$

Voglio una prior che sintetizza sul parametro prima di fare l'esperimento, ma utilizzo informazione dall'esperimento.
($\Rightarrow$ non è coerente con il modo di procedere Bayesiano).

4) Ricorso a $m(\underline{x})$:
Scelgo gli iperparametri facendo ricorso a $m(\underline{x})$ predittiva iniziale.
Ho a disposizione dei campioni passati, ottenuti in passate rilevazioni sullo stesso oggetto, argomento.
Conosco numericamente alcuni momenti marginali di $X \Leftarrow E^m(X)$ valore atteso marginale di X oppure $VAR^m(X)$ varianza marginale di X che sono quantità note.

Ho per $X \rightarrow f(X; \theta)$ in ambito classico e $f(X|\theta)$ e $m(X)$ in ambito Bayesiano.
Considerando la marginale, $E^m(X) = \mu_m$ e $VAR^m(X) = \sigma_m^2$ sono valori noti

che vengono confrontati con i valori che dipendono da θ, poichè provengono dalla legge condizionata $f(x|\theta)$. Questi valori sono: $E^f(X) = \mu_f(\theta)$ e $VAR^f(X) = \sigma_f^2(\theta)$.
Se ho un numero o qualcosa che dipende da θ, ho informazioni su θ.

$$\mu_m = E^\pi(\mu_f(\theta)) = g(\underline{\alpha})$$

Eguaglia un numero ad una funzione degli iperparametri e $\mu_f(\theta)$ è una funzione di $\pi(\theta; \underline{\alpha})$.

$$\mu_m = E^m(X) = \int_{\mathscr{X}} X \cdot m(X)\ dX = \int_{\mathscr{X}} X \int_\Theta f(X|\theta)\pi(\theta; \underline{\alpha})\ dX\ d\theta$$

$$\int_\Theta \pi(\theta; \underline{\alpha}) \int_{\mathscr{X}} X f(X|\theta)\ dX\ d\theta$$

con $\int_{\mathscr{X}} X f(X|\theta)\ dX\ d\theta$ che è uguale a $E^f(X) = E(X|\theta) = \mu_f(\theta)$ e allora:

$$\mu_m = \int_\Theta \mu_f(\theta)\pi(\theta; \underline{\alpha})\ d\theta = E^\pi(\mu_f(\theta))$$

dove $\mu_f(\theta) = h(\theta)$

Se $\underline{\alpha} = (\alpha_1, \alpha_2)$ allora mi serve un'altra equazione: sfrutto le relazioni tra le varianze.

$$VAR^m(X) = \sigma_m^2 = E^\pi(\sigma_f^2(\theta)) + E^\pi[\mu_f(\theta) - \mu_m]^2$$

$E^\pi(\sigma_f^2(\theta)) = \sigma_f^2(\theta)$ corrisponde alla medie delle varianze condizionate ed è uguale a $g_1(\underline{\alpha})$ numero noto.
$E^\pi[\mu_f(\theta) - \mu_m]^2$ corrisponde alla varianza delle medie condizionate che è uguale a $g_2(\underline{\alpha})$ numero noto. Nel caso *Poisson* è θ.

Esercizio
$X|\mu \sim N(\mu, 1) \to N(\mu_0, \sigma_0^2)$
$\mu_m = 1$ e $\sigma_n^2 = 3$

$$\mu_0 \to \mu_f(\mu) = E(X|\mu) = \mu$$

$$\mu_0 \to \mu_m = E^\pi(\mu) = \mu_0 = 1$$

$$\sigma_0^2 \to \sigma_n^2 = E^\pi(\sigma_f^2(\mu)) + E^\pi(\mu - 1)^2 \to 3 = E^\pi(1) + \sigma_0^2$$

dove $E^\pi(1)$ è la varianza del modello e σ_0^2 è la varianza della prior.

5 Sintesi della posterior

Obiettivo fondamentale: sintetizzare la posterior.

$$E[g(\theta)|\underline{x}] = E^{\pi(\cdot|\underline{x})}(g(\theta))$$

Ricordo che

$$E^{\pi(\cdot|\underline{x})} = E^{\pi(\theta|\underline{x})}$$

la media a posteriori di qualunque funzione di θ la calcolo nel seguente modo:

$$\int_{\Theta} g(\theta)\pi(\theta|\underline{x})\ d\theta$$

se $g(\theta)$ fosse θ, allora voglio calcolare la media a posteriori:

$$\begin{cases} \rightarrow\ g(\theta) = \theta \rightarrow\ E(\theta|\underline{x}) \\ \rightarrow\ g(\theta) = I_A(\theta) \rightarrow\ P(\theta \in A|\underline{x}) \end{cases}$$

$I_A(\theta)$ significa che integro su A, cioè su θ che appartiene ad A a posteriori.

Spesso $E^{\pi(\cdot|\underline{x})}(g(\theta))$ **non è calcolabile** perchè:
$\rightarrow$ $m(\underline{x})$ non è calcolabile (quindi la posterior non può essere normalizzata per via analitica).
$\rightarrow$ altre volte non riesco a calcolare analiticamente l'integrale.

Soluzioni:
1) metodi analitici (approssimazione dell'integrale).
2) metodi simulativi.
3) metodi di integrazione numerica.

Evitiamo di parlare del primo poichè è troppo banale ed iniziamo con i metodi analitici.

5.1 Procedure analitiche

Voglio un valore numerico che approssima

$$\int_{\Theta} g(\theta)\pi(\theta|\underline{x})\ d\theta$$

Prendo quindi la posteriori e approssimo con una normale.

5.1.1 Approssimazione normale

Voglio approssimare la posterior $\pi(\theta|\underline{x})$ con una normale. I parametri necessari saranno quindi media e varianza (come capisco se vanno bene?).
Sotto condizioni di regolarità la posterior può essere approssimata con una normale $\sim N(\tilde{\theta}, \tilde{\Sigma})$, dove $\tilde{\theta}$ è il punto di massimo della posterior, cioè la moda della posterior e $\tilde{\Sigma}$ ha come elementi quelli della matrice inversa con il generico elemento j ed è calcolato come segue:

- dalla posterior faccio il logaritmo.
- effettuo poi le derivate seconde.
- confronto con $\tilde{\theta}$.

$$\left[- \frac{\delta^2 \; log \; \pi(\theta|\underline{x})}{\delta\theta_i \; \delta\theta_j} \right]_{\theta=\tilde{\theta}}$$

- faccio l'inversa.

Questo calcolo porterebbe all'informazione osservata se fosse una funzione di verosimiglianza in θ punto di massimo.

$\tilde{\theta}$ corrisponde come già visto alla moda della posterior, cioè il valore di θ che rende massima la posterior:
- voglio massimizzare la posterior.
- valutare le derivate seconde in corrispondenza di $\tilde{\theta}$.

E' possibile calcolare $\tilde{\theta}$ e $\tilde{\Sigma}$ senza conoscere $m(\underline{x})$?
Se la risposta è no, allora mi posso anche fermare.
Se la risposta è si allora:

$$\pi(\theta|\underline{x}) = \frac{f(\underline{x}|\theta)\pi(\theta)}{m(\underline{x})} = f(\underline{x}|\theta)\pi(\theta)$$

dato che:
- il punto che massimizza $\pi(\theta|\underline{x})$ è lo stesso che rende massimo $f(\underline{x}|\theta)\pi(\theta)$ dato che la costante di normalizzazione $m(\underline{x})$ non fa cambiare il punto di massimo.
- se non conosco $m(\underline{x})$ sono in grado di calcolare le derivate seconde, poichè effettuando la trasformazione logaritmica otterrei $-log(m(\underline{x}))$ che non dipende da θ e quindi lo elimino.

Quindi $\tilde{\theta}$ e $\tilde{\Sigma}$ non dipendono da $m(\underline{x})$.

ESERCIZIO - ELEZIONI
$n = 715000$ (numerosità del secondo campione).
$\sum x_i = 350000$ corrisponde al numero di successi (numerosità dei soggetti che votano per centro-sinistra del secondo campione).

→ **Binomiale-Beta**
le informazioni a priori sono le elezioni dell'anno precedente:
$n^* = 690000$ (numerosità del primo campione).
$\sum x_i^* = 390000$ (numerosità dei soggetti che votano per centro-sinistra del primo campione).
Sapendo che $\theta \sim Beta(\alpha, \beta)$ per convenienza.
La posterior sarà allora

$$\theta|\underline{x} \sim Beta(\sum x_i + \alpha, n - \sum x_i + \beta)$$

contenente il numero di successi e insuccessi appartenenti al primo campione, cioè $n^* = 690000$ e $\sum x_i^* = 390000$.
Ottengo quindi

$$\theta|\underline{x} \sim Beta(390000 + \alpha, 300000 + \beta)$$

Sapendo che $\alpha = \sum x_i$ (del secondo campione) $= 350000$ e $\beta = n - \sum x_i$ (secondo campione) $= 365000$, otteniamo:

$$\theta|\underline{x} \sim Beta(\alpha_f = 740000, \beta_f = 665000)$$

Approssimazione con una normale per esempio per calcolare un HPD approssimato (il quale coincide con il credible set della normale).

$$\tilde{\theta} \Rightarrow \theta|\underline{x} \sim Beta$$

ricordo che la moda di una $Beta$ è $\frac{\alpha-1}{\alpha-\beta-1}$ con $\alpha, \beta > 1$.
Prendo α_f e β_f e sostituisco. La moda che si ottiene corrisponde a: 0.5267 punto di massimo della posterior $Beta$.
Per quanto riguarda la varianza della normale, non avremo $\tilde{\Sigma}$ ma avremo uno scalare $\tilde{\sigma}^2$:

$$log\pi(\theta|\underline{x}) = (\alpha_f - 1)log\theta + (\beta_f - 1)log(1-\theta) - log(BETA)$$

$-log(BETA)$ essendo una costante, non la considero nei successivi passi.

$$-\left[-\left(\frac{\alpha_f - 1}{\theta^2}\right) - \left(\frac{\beta_f - 1}{(1-\theta)^2}\right)\right] = \left[\frac{\alpha_f - 1}{\theta^2} + \frac{\beta_f - 1}{(1-\theta)^2}\right]_{\theta=\tilde{\theta}}$$

$$\frac{740000 - 1}{(0.5267)^2} + \frac{665000 - 1}{(1 - 0.5267)^2} = 5636074$$

Il valore ottenuto, corrisponde al reciproco della varianza della normale

$$\tilde{\sigma}^2 = \frac{1}{5636074} = 1.77 \cdot 10^{-7}$$

Ricordo: se fosse stata una matrice, avrei dovuto fare l'inversa.

Posso ora calcolare HPD approssimato: $N(0.5267, 1.77 \cdot 10^{-7})$.

Osservazione:
Con il caso di a priori costanti (prior di Laplace):

$$\pi(\theta|\underline{x}) \propto f(\underline{x}|\theta)$$

poichè la prior $\pi(\theta)$ è una costante.
In questo caso abbiamo che $\tilde{\theta} = \hat{\theta}$ e $\tilde{\Sigma} = \mathscr{I}^{-1}$.

Risultato Bayesiano:	**Risultato Classico:**
$(\theta - \hat{\theta}) \sim N(0, \tilde{\Sigma})$	$(\theta - \hat{\theta}) \sim N(0, \mathscr{I}(\theta, \underline{x}))$
$\theta \rightarrow$ variabile casuale	$\theta \rightarrow$ numero ignoto
$\hat{\theta} \rightarrow$ numero noto	$\hat{\theta} \rightarrow$ variabile casuale

- A livello numerico sono la stessa cosa, danno lo stesso risultato.
- Diverso a livello interpretativo.
- Con LAPLACE, $\tilde{\theta}$ e $\hat{\theta}$ sono la stessa cosa.

Osservazione: quando funziona?
Il risultato può essere applicato anche se non riesco a normalizzare la posterior.
Se $\pi(\theta|\underline{x})$ non è "troppo lontana" dalla normalità.

5.1.2 Approssimazione Laplace

Voglio approssimare

$$E^{\pi(\cdot|\underline{x})}(g(\theta)) = \int_\Theta g(\theta) \cdot \pi(\theta|\underline{x})\ d\theta = \int_\Theta e^{nh(\theta)}\ d\theta$$

approssimazione numerica dell'integrale e non della posterior

$$e^{nh(\theta)} = g(\theta)\pi(\theta|\underline{x}) \rightarrow nh(\theta) = log\{g(\theta)\pi(\theta|\underline{x})\}$$

sviluppo in serie di Taylor di $h(\theta)$ fino al secondo ordine nel punto $\breve{\theta}$ (punto di massimo di $h(\theta)$).

$$\int_\Theta e^{nh(\theta)}\ d\theta \approx \int_\Theta e^{n[h(\breve{\theta}) + (\theta - \breve{\theta})h'(\breve{\theta}) + \frac{(\theta-\breve{\theta})^2}{2}h''(\breve{\theta})]}\ d\theta$$

$$\approx e^{nh(\check{\theta})} \cdot \int_{\Theta} e^{n\frac{(\theta-\check{\theta})^2}{2})h''(\check{\theta})} \, d\theta$$

che può essere riscritta come:

$$\approx e^{nh(\check{\theta})}\sqrt{2\pi}\ \sigma \cdot \int_{\Theta} \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\ e^{n\frac{(\theta-\check{\theta})^2}{2})h''(\check{\theta})} \, d\theta$$

tutto ciò che sta nell'integrale è il nucleo della normale che integra a 1 tale che $\sim N(\check{\theta}, -\frac{1}{nh''(\check{\theta})})$
allora abbiamo

$$e^{nh(\check{\theta})}\sqrt{2\pi}\ \sigma$$

$$e^{nh(\check{\theta})}\sqrt{2\pi}\ \sqrt{-\frac{1}{nh''(\check{\theta})}}$$

$$E^{\pi(\cdot|\underline{x})}(g(\theta)) = \int_{\Theta} e^{nh(\theta)} \, d\theta$$

che corrisponde ad un numero. Al fine di calcolarlo ho bisogno di:

$$\check{\theta}: \ h(\theta) = \frac{1}{n} log[g(\theta)\pi(\theta|\underline{x})]$$

con

$$\pi(\theta|\underline{x}) = \frac{f(\underline{x}|\theta)\pi(\theta)}{m(\underline{x})}$$

ricordo che $m(\underline{x})$ può essere eliminata se non sono in grado di calcolarla, poichè non varia il punto di massimo. Quindi $\check{\theta}$ può essere calcolata anche se $m(\underline{x})$ è ignota.

Con Laplace questo non può essere fatto, poichè in h la $m(\underline{x})$ è presente.
Valutare h in $\check{\theta}$: $h(\check{\theta})$ richiede la conoscenza di $m(\underline{x})$.

$$\int_{\Theta} g(\theta)\pi(\theta|\underline{x}) \, d\theta = \int_{\Theta} g(\theta)\frac{f(\underline{x}|\theta)\pi(\theta)}{m(\underline{x})} \, d\theta = \frac{\int_{\Theta} g(\theta) f(\underline{x}|\theta)\pi(\theta) \, d\theta}{\int_{\Theta} f(\underline{x}|\theta)\pi(\theta) \, d\theta}$$

ho potuto fare questo poichè $m(\underline{x})$ è una costante e quindi posso portarla fuori dall'integrale. Una volta fuori dall'integrale, so che assume quel determinato valore.
Facendo l'approssimazione di Laplace al numeratore e al denominatore ottengo:

$$\frac{e^{nh(\theta)}}{e^{nh^*(\theta)}}$$

5.2 Metodi simulativi

I metodi di simulazione ricostruiscono le caratteristiche di una distribuzione complessa (nel nostro caso la posterior) tramite un campione di elementi indipendenti generato dalla distribuzione.

5.2.1 Metodo Monte Carlo

Siamo interessati alla posterior $\pi(\theta|\underline{x})$ (o conosco il nucleo ma non riesco a normalizzare oppure la conosco ma non riesco a calcolare $E^{\pi(\cdot|\underline{x})}[g(\theta)]$, quindi o non conosco $m(\underline{x})$, oppure lo conosco ma non sono in grado di calcolare $E^{\pi(\cdot|\underline{x})}[g(\theta)]$).
Ho quindi a disposizione una successione di realizzazioni indipendenti e identicamente distribuiti $\theta_1, \theta_2, ..., \theta_m$.
Prima di tutto calcolo $g(\theta_1), g(\theta_2), ..., g(\theta_m)$ da cui posso poi ricavarne la media campionaria $\bar{g}_m = \frac{1}{m}\sum_{i=1}^{m} g(\theta_i)$, questo è stimatore non distorto, rispetta la legge forte dei grandi numeri tale per cui

$$P(\bar{g}_m \to E^{\pi(\cdot|\underline{x})}[g(\theta)]) = 1$$

e vale il Teorema Centrale del Limite se

$$\sigma^2 = \int_{\Theta} [g(\theta) - E^{\pi(\cdot|\underline{x})} g(\theta)]^2 \pi(\theta|\underline{x})\, d\theta < \infty$$

cioè

$$\sqrt{m}(\bar{g}_m - E^{\pi(\cdot|\underline{x})} g(\theta)) \xrightarrow{d} N(0, \sigma^2)$$

Questo risulta utile per calcolare gli intervalli di confidenza:

$$[\bar{g}_m \pm Z_{1-\frac{\alpha}{2}} \sqrt{\hat{Var}(\bar{g}_m)}]$$

Per calcolare $\hat{Var}(\bar{g}_m)$ parto dal valore reale $Var(\bar{g}_m) = \frac{1}{m}\sum_{i=1}^{m} g(\theta_i) = \frac{\sigma^2}{m}$ con $\sigma^2 = VAR^{\pi(\cdot|\underline{x})}[g(\theta)]$ e poichè so che

$$\hat{\sigma^2} = \frac{1}{m-1}\sum_{i=1}^{m}(g(\theta_i) - \bar{g}_m)^2 \to \hat{Var}(\bar{g}_m) = \frac{\sum_{i=1}^{m}(g(\theta_i) - \bar{g}_m)^2}{m(m-1)}$$

sapendo che $\hat{\sigma}^2$ è la stima di Monte Carlo.

5.2.2 MCIS (Monte Carlo Importance Sampling)

In pratica vengono aggiunte funzioni ausiliarie per trovare un campione indipendente e identicamente distribuito della posterior.

Prendo la funzione $h(\theta)$ da cui deve essere facile estrarre determinazioni casuali e deve essere sufficientemente "vicina" a $\pi(\theta|\underline{x})$.

Algoritmo:
1) Estrarre un campione indipendente e identicamente distribuito da $h(\theta) \to \theta_1', \theta_2', ..., \theta_m'$, questo non deriva dalla posterior ma deve avere lo stesso supporto. Ora mi servono dei pesi che mi bilanciano h (quando è vicina alla posterior i pesi saranno 1 mentre quando è lontana la deve riavvicinare).
2) Voglio approssimare per via simulativa

$$E^{\pi(\cdot|\underline{x})}[g(\theta)] \to \tilde{g}_m = \frac{1}{m}\sum_{i=1}^{m} \omega_i g(\theta_i')$$

dove

$$\omega_i = \frac{\pi(\theta_i'|\underline{x})}{h(\theta_i')}$$

e

$$\theta_i' : \ \frac{h(\theta_i')}{w_i}$$

Con $h(\theta_i') > \pi(\theta_i'|\underline{x})$ e $w_i < 1$ (con il metodo Monte Carlo ho $\omega_i = 1$ e $\theta_i' = \theta_i$).
Ma come calcolo $\pi(\theta_i'|\underline{x})$?
Mi interessa

$$E^{\pi(.|\underline{x})}[g(\theta)] = \int_\Theta g(\theta)\pi(\theta|\underline{x})\, d\theta$$

posso moltiplicare per $\frac{h(\theta)}{h(\theta)}$

$$\int_\Theta \frac{g(\theta)\pi(\theta|\underline{x})}{h(\theta)} h(\theta)\, d\theta = \int_\Theta g^*(\theta)h(\theta)\, d\theta = E^{h(\cdot)}[g^*(\theta)]$$

Non conosco la legge di distribuzione di $g^*(\theta)$.

Quindi ho $\theta_1', \theta_2', ..., \theta_m'$ da $h(\theta)$ e applico Monte Carlo a

$$E^{h(\cdot)}[g^*(\theta)] \to g^*(\theta_1'), g^*(\theta_2'), ..., g^*(\theta_m')$$

con $g^*(\theta_1') = \frac{g(\theta_1')\cdot\pi(\theta|\underline{x})}{h(\theta_1')}$ ecc... e quindi

$$\tilde{g}_m = \frac{1}{m}\sum_{i=1}^{m} g^*(\theta_i') = \frac{1}{m}\sum_{i=1}^{m} \frac{\pi(\theta_i'|\underline{x})}{h(\theta)} g(\theta_i')$$

Osservazione: se $\pi(\cdot|\underline{x})$ non è nota è un problema, devo fare un altro passo (se fosse stata nota, mi sarei fermato).

$$E^{\pi(.|\underline{x})}g(\theta) = \int_\Theta g(\theta)\pi(\theta|\underline{x})\, d\theta = \int_\Theta g(\theta)\frac{\pi(\theta)f(\underline{x}|\theta)}{m(\underline{x})}\, d\theta = \frac{\int_\Theta g(\theta)\pi(\theta)f(\underline{x}|\theta)\, d\theta}{\int_\Theta \pi(\theta)f(\underline{x}|\theta)}$$

se $\pi(\theta)$ è propria posso farla diventare funzione di importanza $h(\theta)$

$$= \frac{\int_\Theta g^*(\theta)\pi(\theta)\,d\theta}{\int_\Theta g^{**}(\theta)\pi(\theta)\,d\theta}$$

Quindi la stima è:

$$\frac{\frac{1}{m}\sum i g(\theta_i')f(\underline{x}|\theta')}{\frac{1}{m}\sum i f(\underline{x}|\theta')}$$

Osservazione: anche per MCIS vale la legge forte dei grandi numeri e il teorema centrale del limite.

5.2.3 Metodo MCMC

RIPASSO (catene di Markov): $\{X^n, n \in \mathbb{N}\}$ è un processo a tempo discreto con la proprietà della perdita di memoria (markovianità)

$$P(X^{n+1} \in A|X^n = x_n, ..., X^0 = x_0) = P(X^{n+1} \in A|X^n = x_n)$$

$$A \subseteq S \text{ spazio stato}$$

· **Nucleo di transizione:** $P(x, A) = P(X^{n+1} \in A|X^n = x)$
· La catena è omogenea se $P(x, A)$ non dipende da n.
· Ho bisogno della distribuzione iniziale π, la marginale di X^0.
· Kernel di transizione in m passi $\rightarrow P^m(x, A) = P(X^m \in A|X^0 = x)$
· π è stazionario se è la distribuzione marginale di ogni elemento X^n

L'obbiettivo è trovare una catena di Markov omogenea con distribuzione iniziale stazionaria (e propria) che coincide con la posterior ($\pi(\theta|\underline{x})$).

Teorema ergodicità: sia $\{X^n\}$ una catena di Markov omogenea, irriducibile, ricorrente e con nucleo di transizione P e distribuzione stazionaria π. Allora:
1) π è l'unica distribuzione stazionaria
2) Sia g una funzione a valori reali tale che $E^\pi[g(X)] < \infty$ e sia $x^1, ..., x^m$ una realizzazione finita della catena con $\bar{g}_m = \frac{1}{m}\sum_{i=1}^m g(x^i)$, allora

$$P(\bar{g}_m \xrightarrow[n\to\infty]{} E^\pi[g(X)]) = 1$$

3) Se la catena è aperiodica allora

$$dist(P^n(x, .) - \pi(.)) \xrightarrow[n\to\infty]{} 0$$

cioè ad un certo punto estrarre dalla condizionata o dalla marginale non cambia perchè la distanza tra esse è 0.

Osservazione: da un certo punto in poi le realizzazioni x_i possono ritenersi provenire da π (che sarà la nostra posterior) $\rightarrow$ **Burn-in period**. Inoltre il secondo punto ci permette di applicare Monte Carlo anche se le osservazioni non sono indipendenti.

$$x^0, ..., x^m \text{ (Burn-in period) } x^{m+1}, ..., x^{m+n} \text{ (ampiezza campione MC)}$$

Posso pensare che dall'osservazione $m - esima$ le realizzazioni vengono dalla posterior e dunque ne estraggo altre n per applicare Monte Carlo.
$\{\theta^n, n \in \mathbb{N}\}$ con distribuzione iniziale stazionaria $\pi(\theta|\underline{x}) = \theta^0$
L'idea è inizializzare θ^0, si lascia decorrere un certo periodo di tempo per la stabilizzazione e poi si pensa che le osservazioni vengono dalla posterior. Quindi voglio che θ^0 sia determinazione della variabile casuale Θ^0 che si distribuisce come la posterior.

Ma come inizializzo θ^0 dalla posterior se questa non è nota?
Si assegna θ^0 casuale, si potrebbe prendere dalla prior se questa é propria (il burning period deve comunque inglobare le realizzazioni necessarie perchè θ' derivi dalla posterior).

Ci sono due diverse tecniche: Gibbs Sampling e Metropolis Hasting.

Gibbs Sampling

La tecnica di **Gibbs Sampling** si usa quando il parametro ha tante componenti $\underline{\theta} = (\theta_1, ...\theta_d)$, dove la generica θ_i non è necessariamente unidimensionale; si suppone di conoscere (e di poter estrarre da) la legge di distribuzione di $\pi(\theta_i|\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_d)$, che prende il nome di **full conditional**.

Algoritmo per la generazione delle full conditionals:
1) Inizializzare la catena in $\underline{\theta}^0 = (\theta_1^0, ..., \theta_d d^0)$
2) Si ottiene l'aggiornamento $\underline{\theta}^1 = (\theta_1^1, ..., \theta_d^1)$ dove il generico θ_i^1 é estratto da $\pi(\theta_i|\theta_1^1, ..., \theta_{i-1}^1, \theta_{i+1}^0, ..., \theta_d^0)$
3) Si reitera il passo precedente con i dovuti aggiornamenti.

ESEMPIO (Two stage Gibbs Sampling)
Dati $x|\theta \sim Bi(n, \theta)$ e $\theta \sim Beta(\alpha, \beta)$, si può implementare Gibbs Sampling per ottenere un campione da $dist(x, \theta)$ che sarebbe $distribuzione(x, \theta)$?

Ho le **full conditionals** $\rightarrow (x, \theta)$ e le **gibbs sampling** $\rightarrow (x|\theta)$ e $(\theta|x)$ tali che:

$$x|\theta \sim Binomiale(n, \theta)$$

$$\theta|x \sim Beta(\alpha + x, \beta + n - x)$$

Visto che conosco le due distribuzioni posso applicare la tecnica di Gibbs Sampling:
1) Trovo x^0 e θ^0, ad esempio visto che $\theta \sim Beta$ prendo dei valori tra 0 e 1 oppure un'estrazione da una $Beta(\alpha, \beta)$
So che x^0 è un intero tra 0 e 1 oppure estrazioni da $binomiale(n, \theta^0)$.
2) Ora calcolo $(x^1, \theta^1) : x^1 \rightarrow Bi(n, \theta^0)$ é la legge di distribuzione di $X|\theta = \theta^0$.
Invece $\theta^1 \rightarrow Beta(\alpha + x^1, \beta + n - x^1)$.
Passo $(x^2, \theta^2) : x^2 \rightarrow Bi(n, \theta^1)$ per $X|\theta = \theta^1$,e $\theta^1 \rightarrow Beta(\alpha + x^2, \beta + n - x^2)$
Una volta ottenuti $(x^0, \theta^0), ..., (x^m, \theta^m)$ e raggiunto il burn-in period, butto via quanto trovato e con le realizzazioni successive applico Monte Carlo.
Alternativamente potrei inizializzare m catene in parallelo:

$$\begin{cases} (x^0, \theta^0) \\ (x^0, \theta^0) \\ ... \\ (x^0, \theta^0) \end{cases}$$

da cui ottengo n osservazioni che genera il campione Monte Carlo:

$$\begin{cases} (x^{m+1}, \theta^{m+1}) \\ (x^{m+1}, \theta^{m+1}) \\ ... \\ (x^{m+1}, \theta^{m+1}) \end{cases}$$

che costituiscono il campione.
Con il primo metodo le osservazioni del campione Monte Carlo non sono indipendenti, con il secondo metodo lo sono.

Esercizio
Con la Normale abbiamo visto che se ne conosco la media ($X \sim N(\mu = \text{nota}, \sigma^2)$) la coniugata è una $GaInv(\alpha, \beta)$, se conosco la varianza σ^2 ($X \sim N(\mu, \sigma^2 = \text{nota})$) la coniugata è $N(\mu_0, \sigma_0^2)$.

Gibbs-Sampling: date $\mu \sim N(0, 1)$ e $\sigma^2 \sim GaInv(1, 1)$, quali sono le full conditionals?

Ipotizzo che μ e σ^2 sono indipendenti e scrivo la coniugata (numeratore della posterior coniugata)

$$(\mu, \sigma^2, \underline{x}) \to \Psi(\mu, \sigma^2, \underline{x}) = f(\underline{x}|\mu, \sigma^2)\pi(\mu)\pi(\sigma^2) =$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{\mu^2}{2}} \cdot \frac{1}{(\sigma^2)^2} e^{-\frac{1}{\sigma^2}}$$

So che $m(\underline{x}) = \int_{\mathbb{R}} \int_{\mathbb{R}^+} \Psi(\mu, \sigma^2, \underline{x})\, d\mu\, d\sigma^2 \to$ non ha soluzione analitica, quindi applico Gibbs Sampling.

Faccio le full conditional:

$$\pi(\mu|\sigma^2, \underline{x}) = \frac{\pi(\mu, \sigma^2|\underline{x})m(\underline{x})}{\pi(\sigma^2|\underline{x})m(\underline{x})} =$$

$$= \frac{\Psi(\mu, \sigma^2, \underline{x})}{\tilde{\Psi}(\sigma^2, \underline{x})} \propto \Psi(\mu, \sigma^2, \underline{x})$$

uso la proporzionalità perchè mi interessano solo gli elementi che dipendono da μ, il denominatore è una costante.
Allora

$$\pi(\mu|\sigma^2 = \text{nota}, \underline{x}) \propto e^{-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2} e^{-\frac{\mu^2}{2}}$$

sfruttando il risultato noto sulla famiglia coniugata

$$\Rightarrow N(\frac{\sum_i x_i}{\sigma^2 + n}, \frac{\sigma^2}{\sigma^2 + n})$$

Per la seconda full conditional:

$$\pi(\sigma^2|\mu, \underline{x}) \propto \frac{1}{(\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2} \frac{1}{(\sigma^2)^2} e^{-\frac{1}{\sigma^2}}$$

Vedo che il risultato è ancora sulla famiglia esponenziale coniugata per σ^2:

$$\Rightarrow GaInv(\frac{n}{2} + 1, \frac{1}{2}\sum_i (x_i - \mu)^2 + 1)$$

con $\alpha = 1$ e $\beta = 1$.

Quando mi viene data $X \sim N(\mu, \sigma^2)$ con i due parametri ignoti e calcolo le full conditional noto che:
1) il Gibbs Sampling da luogo ad una catena di Markov (questa dipende solo dal passo precedente).
2) la catena è omogenea (cioè non dipende dal passo in cui ci troviamo),

infatti le full conditional sono sempre le stesse.
3) il nucleo di probabilità di transizione è ottenuto dal prodotto delle full conditionals:

$$P(\theta^j, \theta^{j+1}) = \prod_{i=1}^{d} \pi(\theta_i | \theta_1^{j+1}, \theta_{2^{j+1}}, ..., \theta_{i-1}^{j+1}, \theta_{i+1}^{j}, ..., \theta_d ? j, \underline{x})$$

4) la catena è aperiodica, irriducibile e ricorrente $\Rightarrow$ posso applicare il teorema dell'ergodicità.

Metropolis Hastings

La seconda tecnica per il metodo MCMC è il **Metropolis Hastings** e si tratta di un algoritmo di accettazione-rifiuto. Indico con $q(\theta, \theta')$ la legge di θ' condizionata a θ, in particolare è la **legge di distribuzione proposta**, cioè quella che fornisce i candidati per il passo successivo a quello in cui mi trovo.
Estraggo da q un candidato θ' e la sua probabilità di accettazione $\alpha(\theta, \theta')$ è pari a:

$$\begin{cases} min\{\frac{\pi(\theta'|\underline{x})q(\theta',\theta)}{\pi(\theta|\underline{x})q(\theta,\theta')}, 1\} & \text{se } \pi(\theta|\underline{x})q(\theta,\theta') \neq 0 \\ 1 & \text{se } \pi(\theta|\underline{x})q(\theta,\theta') = 0 \end{cases}$$

Algoritmo:
1) si fissa (o inizializza) θ^0 scegliendo arbitrariamente (posso usare la media, la mediana, la media condizionata,...).
2) si genera θ' dalla proposal $q(\theta^0, \theta')$ e a questo punto deve decidere se accettare o meno il candidato θ'.
3) l'obbiettivo è decidere se accettare θ' utilizzando la probabilità $\alpha(\theta^0, \theta') \rightarrow$ estraggo u da $U(0,1)$ e se $u \leq \alpha(\theta^0, \theta')$ accetto il candidato e quindi $\theta^1 = \theta'$, altrimenti $\theta^1 = \theta^0$.
4) ripeto i passi 2 e 3 fino alla lunghezza che desidero per la catena

Caratteristiche Metropolis:
- $\alpha(\theta, \theta')$ prescinde dalla conoscenza della costante di normalizzazione della posterior $m(\underline{x})$.
- la catena che realizzo è reversibile.

DEFINIZIONE: si parla di **reversibilità** per una catena di Markov quando $P(X^{n-1} = x, X^n = y) = P(X^{n-1} = y, X^n = x)$.
Allora $\pi(x)P(x,y) = \pi(y)P(y,x)$

Ora analizzo il nucleo di transizione nel Metropolis:

$$q(\theta, \theta^{'})\alpha(\theta, \theta^{'}) = P(\theta, \theta^{i})$$

Se la proposal rendesse la catena reversibile avrei:

$$\pi(\theta|\underline{x})q(\theta, \theta^{'}) = \pi(\theta^{'}|\underline{x})q(\theta^{'}, \theta) \Rightarrow \alpha(\theta, \theta^{'}) = 1$$

cioè ogni candidato viene accettato.
Non esiste una q che mi dia questi risultati ma:

$$\exists(\theta, \theta^{'}) : \pi(\theta|\underline{x})q(\theta, \theta^{i}) > \pi(\theta^{'}|\underline{x})q(\theta^{'}, \theta)$$

Quindi il numero dei passaggi da θ a $\theta^{'}$ (cioè la parte a sinistra della disequazione) è troppo grande per avere reversibilità ed è troppo piccolo il numero di passaggi da $\theta^{'}$ a θ (la parte a destra). Le probabilità di accettazione sono $\alpha(\theta, \theta^{'}) < 1$ e $\alpha(\theta^{'}, \theta) = 1 \rightarrow$ la probabilità di passaggio da θ a $\theta^{'}$ viene fatta minore di 1 mentre è massimizzata quella di passaggio da $\theta^{'}$ a θ.

Metropolis Random Walk
Prendo $\theta^{'} = \theta + x$ dove θ è fissato e x è una perturbazione aleatoria, in questo modo ogni candidato è ancorato alla catena al tempo precedente (tipicamente per la x si prende una Normale, una T o una Uniforme).
Se $X \sim U$, $q(\theta, \theta^{'})$ è semplicemente una uniforme traslata.
Poichè

$$\alpha(\theta, \theta^{'}) = \frac{\pi(\theta^{'}|\underline{x})q(\theta^{'}, \theta)}{\pi(\theta|\underline{x})q(\theta, \theta^{'})}$$

mi servirebbe la posterior $\rightarrow$ prendo il prodotto tra likelihood e prior.

Indipendent Metropolis
$\theta^{'}$ è estratto da qualche legge $h(\theta^{'})$ e dunque il candidato $\theta^{'}$ non dipende dal candidato al tempo precedente.

$$\alpha(\theta, \theta^{'}) = \frac{\pi(\theta^{'}|\underline{x})q(\theta^{'}, \theta)}{\pi(\theta|\underline{x})q(\theta, \theta^{'})}$$

L'estrazione è indipendente ma $\alpha(\theta, \theta^{'})$ dipende comunque da θ, dunque le realizzazioni non sono indipendenti.

Osservazioni:
1) Il tasso di accettazione è ottimale tra il 30% e il 70%. Se è molto basso ho

inefficienza algoritmica, tipicamente vuol dire che mi soffermo dove la posterior ha densità più bassa e non riesco ad prendere osservazioni dalle zone in cui la densità è più elevata. Se è molto alta i valori vengono tipicamente dalle zone con densità maggiore della posterior, però facendo un campionamento di Monte Carlo la versione empirica della posterior sarà concentrata solo sulle zone con elevata densità; devo fare tuning sulla proposal per cercare di prendere anche i punti con densità più bassa.
2) Come scelgo l'algoritmo? Se $\underline{\theta}$ è multidimensionale ed ha tante componenti provo ad usare Gibbs. Se $\underline{\theta} = (\theta_1, ..., \theta_5)$ e non riesco a calcolare una full conditional, posso implementare un Metropolis within Gibbs (applico un Metropolis alla full conditional che non conosco).
3) m=lunghezza del burning period, come scelgo m? Come scelgo n?
Uso strumenti di diagnostica della convergenza.Ad esempio estraggo un campione di 100 osservazioni e faccio l'istogramma di $\underline{\theta}$, se passo ad esempio a $m + 10$ ed estraggo un campione facendone l'istogramma posso confrontare i due grafici per capire se si è raggiunta la stabilizzazione. Oppure posso analizzare le medie delle realizzazioni ad ogni passo e vedo quando queste si stabilizzano. Questa procedura però è rischiosa anche perchè la posterior non è nota e non esiste un m ottimale, gli strumenti di diagnostica non verificano se si viene dalla posterior ma se i risultati si stabilizzano.
4) Se voglio un'ampiezza del campione Monte Carlo pari a n posso:
- Inizializzo una catena fino a raggiungere il passo $m-1$, dal passo successivo prendo n realizzazioni.
- Inizializzo n catene, una volta raggiunto il passo $m - 1$ prendo la realizzazione successiva di ogni catena (quindi ho n osservazioni).

Con il primo metodo devo fare $m + n$ estrazioni, con il secondo ne faccio $m \cdot n$. Le estrazioni del primo metodo però sono dipendenti, il secondo fornisce stime meno distorte.
Una buon compromesso è inizializzare una catena, una volta arrivati al passo $m - 1$ scelgo una osservazioni ogni k (thinning period); in questo modo devo estrarre $m + n \cdot k$ osservazioni.

6 Approccio decisionale

Come si fa inferenza bayesiana?

$$\pi(\theta|\underline{x}) \rightarrow \text{ calcolata}$$

L'ho calcolata in qualche modo, ho quindi a disposizione tutte le informazioni necessarie sulla posterior.
Posso pensare di fare qualcosa di più affidabile, di più teorico.

Approccio decisionale in inferenza bayesiana (potrei utilizzare questo approccio anche in ambito classico).
C'è quindi un decisore, colui che prende le decisioni che può essere un soggetto qualunque. Il decisore deve scegliere un'azione $\mathbf{a} \in \mathcal{A}$.
$\mathcal{A}$ è la classe delle azioni.

Le **conseguenze** di ogni **a** dipendono da:
$\rightarrow$ qualcosa non dominata da me.
$\rightarrow$ da θ "stato di natura" $\in \Theta$ spazio dei possibili stati di natura.

Una generica conseguenza dipende da **a** e dai possibili valori di θ.

$$C_a(\theta) \Leftarrow \textbf{ si suppone che siano totalmente ordinate}$$

Totalmente ordinate significa che su 2 possibili conseguenze so dire quale preferisco o se sono indifferente.
Inoltre so che $C_a(\theta) \in \varphi$
Voglio trasformare queste conseguenze in numeri: **le trasformo in perdite**
Supponiamo che esista una $f : \varphi \rightarrow \mathbb{R}$ che mi permette di generare dalla conseguenza un numero reale.
Allora la funzione di perdita f la scriviamo come $L =$ Loss ed è composta da 2 argomenti: $L(\theta, a)$

$L(\theta, a)$ è la perdita che il decisore consegue se sceglie **a** quando lo stato di natura è θ.
Voglio minimizzare, quindi devo decidere tra 2 funzioni di perdita quale preferisco. Molto spesso ci si trova in una condizione in cui il decisore non sa quale azione è preferita all'altra. Viene quindi aggiunto un terzo elemento alla funzione di perdita: **il criterio di ottimalità: K**

$$\mathcal{L} = \{L(\theta, a) : \ a \in \mathcal{A}\}$$
$$K : \mathcal{L} \rightarrow \mathbb{R}$$

cioè ad ogni funzione di perdita (L) associo un numero.

$\Rightarrow \tilde{a}$: **azione ottima** (preferita con il criterio di ottimalità K) è:

$$K[L(\theta, \tilde{a}] < K[L(\theta, a) \ \forall \ a \neq \tilde{a}]$$

Voglio quella che mi da la sintesi più bassa.

Genero quindi la quaterna:

$$(\Theta, \mathcal{A}, L(\theta, a), K)$$

Θ (spazio degli stati di natura), $\mathcal{A}$ (insieme delle azioni), $L(\theta, a)$ (funzione di perdita) e K (criteri di ottimalità).
Forma canonica di un problema decisionale in condizioni di incertezza.

Vediamo le caratteristiche di $\mathcal{A}$:
$\mathcal{A} \Leftarrow \ a_1 \succcurlyeq a_2$ relazione di preferenza debole (binaria). In generale per scegliere tra 2 azioni che fanno un ordinamento.

$$a_1 \succcurlyeq a_2 \Leftrightarrow L(\theta, a_1) \leq L(\theta, a_2) \ \forall \theta \in \Theta$$

E' un preodinamento parziale:
Preordinamento $\rightarrow$ vale se valgono le seguenti proprietà:
- Riflessiva: $a \succcurlyeq a$.
- Transitiva: $a_1 \succcurlyeq a_2$ e $a_2 \succcurlyeq a_3$, allora $a_1 \succcurlyeq a_3$.
- [non gode della proprietà antisimmetrica: $a_1 \succcurlyeq a_2$ e $a_2 \succcurlyeq a_1 \rightarrow a_1 = a_2$]

Parziale $\rightarrow$ non tutte le coppie possono essere messe in relazione.
$\exists$ coppie di azioni non confrontabili sotto questo profilo di relazione binaria.

6.1 Ammissibilità

Voglio ridurre la dimensione di $\mathcal{A} \rightarrow$ **"Ammissibilità"**.

$\rightarrow$ Un'azione è **ammissibile** se $\nexists$ un'altra azione (migliore)

$$a^{'} \in \mathcal{A} \text{ tale che } a^{'} \succ a$$

$a^{'} \succ a$ significa che esiste almeno un punto dove è minore stretto

$$a^{'} \succ a \text{ sse } L(\theta, a^{'}) \leq L(\theta, a) \ \forall \theta \in \Theta$$

e

$$\exists \theta \in \Theta \text{ tale che } L(\theta, a^{'}) < L(\theta, a)$$

Esiste un preordinamento? Gode delle due proprietà?

→ Un'azione è **inammissibile** se ∃ un'azione (migliore)

$$a^{'} \in \mathcal{A} \text{ tale che } a^{'} \succ a$$

Esempio

	a_1	a_2	a_3
θ_1	1	3	4
θ_2	-1	5	5
θ_3	0	-1	-1

$\Theta = \{\theta_1, \theta_2, \theta_3\}$
$\mathcal{A} = \{a_1, a_2, a_3\}$
$L(\theta_1, a_1), L(\theta_2, a_2), L(\theta_3, a_3)$ assumeranno 3 valori → sfrutto la matrice
a_1 e a_2 non sono confrontabili
tra a_2 e a_3 c'è una relazione
$a_2 \succ a_3$ poichè

$$[L(\theta, a_2) \leq L(\theta, a_3)\ \forall \theta \in \Theta] \text{ e } L(\theta, a_2) < L(\theta, a_3)$$

Quindi a_2 rende a_3 **inammissibile**.

E' opportuno restringere l'attenzione all'insieme $\mathcal{A}^+$ (delle azioni ammissibili).
$\mathcal{A}^+ \subseteq \mathcal{A}$ a questo punto devo decidere un certo criterio di ottimalità

$$K : \mathcal{L} \to \mathbb{R}$$

Criterio Bayesiano ⇒ θ è una variabile casuale.
Criterio non Bayesiano ⇒ θ non è una variabile casuale, ma un parametro ignoto.

6.1.1 Criteri Bayesiani

$K[L(\theta, a)]$ = numero con θ variabile casuale e $L(\theta, a)$ che varia al variare di θ ed è anche essa una variabile casuale (y) con una certa legge.

1) **Criterio** (K) **del valore atteso:**
$\theta \to \pi(\theta)$ (che corrisponde alla legge di θ e non per forza la prior).

$$K[L(\theta, a)] = E^{\pi}[L(\theta, a)] = \int_{\Theta} L(\theta, a) \cdot \pi(\theta)\ d\theta$$

nel caso discreto utilizzo la sommatoria $\sum$.
Mi accontento quindi di ciò che la funzione di **a** mi fa perdere **in media**.

2) Criterio media-varianza:

$$K[L(\theta, a)] = E^{\pi}[L(\theta, a)] + \alpha \cdot Var^{\pi}[L(\theta, a)]$$

Abbiamo $\alpha > 0$ e prendiamo la varianza con il variare di α.
Questo metodo ha dei problemi, ad esempio non posso confrontare media e varianza, poichè sono differenti.

3) Criterio della soglia critica:
Mi preoccupo solo delle soglie grandi, cioè dei θ che rendono alta la perdita. Voglio ridurre la probabilità di avere perdite elevate, quindi probabilizzo i θ che stanno sopra la soglia $\lambda > 0$ dove è grande $L(\theta, a)$.

$$K(L(\theta, a)) = P\{\theta : L(\theta, a) > \lambda\}$$

6.1.2 Criteri non Bayesiani

In ambito classico non si adotta un approccio decisionale generalmente.
θ è un parametro fisso.

1) MiniMax:

$$K[L(\theta, a)] = \sup_{\theta \in \Theta} L(\theta, a)$$

Considerando una rappresentazione grafica in cui ho $L(\theta, a^{'})$ simile ad una normale ed una seconda funzione di perdita $L(\theta, a^{''})$ costante son punto di massimo inferiore alla prima funzione di perdita, allora ho che $a^{''}$ è ottimale perchè sto valutando solamente il sup.

Perchè utilizziamo solo il criterio del valore atteso?

Monotonia dei K:
Ho scoperto che

$$a_1 \geq a_2 \ [\Leftrightarrow L(\theta, a_1) \leq L(\theta, a_2) \ \forall \theta \in \Theta]$$

sintetizzo con lo stesso criterio (funzione di perdita K)

$$K[L(\theta, a_1)] \leq K[L(\theta, a_2)]$$

se succede questo, allora K è monotono.

Questo vale solo per il criterio del valore atteso mentre per il criterio Media-varianza non è monotono.

Esempio

	a_1	a_2
θ_1	2	3
θ_2	0	3

$$L(\theta, a_1) = \begin{cases} 2 & \text{se} \quad \theta = \theta_1 \\ 0 & \text{se} \quad \theta = \theta_2 \end{cases}$$

$$L(\theta, a_2) = \begin{cases} 3 & \text{se} \quad \theta = \theta_1 \\ 3 & \text{se} \quad \theta = \theta_2 \end{cases}$$

Qualunque sia la legge di π, vedendo che i valori di $a_2 \geq a_1$ sempre, allora:

$$E^{\pi}[L(\theta, a_1)] \leq E^{\pi}[L(\theta, a_2)]$$

quindi a_2 è inammissibile

Il criterio media-varianza riesce però a rendere ottima un'azione inammissibile.
$L(\theta, a_2)$ assume valore 3 con probabilità 1 e $Var^{\pi}[L(\theta, a_2)] = 0$.

N.b. Con α piccolo vince a_1, ma se alzo α, la penalità incide tanto e rischio quindi di superarla (media-varianza).

$$?a : \tilde{a} = a_2$$

Il criterio media-varianza non è monotono.

6.1.3 Funzioni di perdita

1) Perdita quadratica:

$$L(\theta, a) = (\theta - a)^2 \text{ si comporta in modo simmetrico}$$

$$L(\theta, a) = w(\theta - a)^2 \text{ ponderate con } w \text{ funzione di peso}$$

2) Perdita lineare

$$L(\theta,a)=|\theta-a| \textbf{ assoluta} = \begin{cases} \theta-a & \theta \geq a \quad a \text{ sottostima } \theta \\ a-\theta & \theta < a \quad a \text{ sovrastima } \theta \end{cases}$$

$$L(\theta,a) \textbf{ lineare} = \begin{cases} K_1(\theta-a) & \text{se} \quad \theta \geq a \\ K_2(a-\theta) & \text{se} \quad \theta < a \end{cases}$$

$$L(\theta,a) \textbf{ pesata} = \begin{cases} K_1(\theta)(\theta-a) & \theta \geq a \quad a \text{ sottostima } \theta \\ K_2(\theta)(\theta-a) & \theta < a \quad a \text{ sovrastima } \theta \end{cases}$$

3) Perdita 0/1

$$\text{Stima: } L(\theta,a) = \begin{cases} 0 & |\theta-a| \leq \varepsilon \quad \varepsilon > 0 \\ 1 & |\theta-a| > \varepsilon \end{cases}$$

[non perdo niente se l'errore di stima è piccolo]

6.1.4 Verifica d'ipotesi

$$H_0 : \theta \in \Theta_0 \qquad H_1 : \theta \in \Theta_1$$

$\mathcal{A} = \{a_0, a_1\}$ con a_0 (accetto l'ipotesi nulla) e a_1 (rifiuto l'ipotesi nulla).

$$L(\theta,a_0) = \begin{cases} 0 & \text{se } \theta \in \Theta_0 \\ 1 & \text{se } \theta \in \Theta_1 \end{cases}$$

$$L(\theta,a_1) = \begin{cases} 0 & \text{se } \theta \in \Theta_1 \\ 1 & \text{se } \theta \in \Theta_0 \end{cases}$$

Quanto perdo in questi casi? Ho bisogno di 2 funzioni.
Uso perdita 0/1

$$L(\theta,a_i) = \begin{cases} 0 & \text{se } \theta \in \Theta_i \\ 1 & \text{se } \theta \in \Theta_j \end{cases}$$

Con $i = 0, 1$ e $j = 0, 1$.

$$L(\theta,a_i) = \begin{cases} 0 & \text{se } \theta \in \Theta_i \\ K_i & \text{se } \theta \in \Theta_j \end{cases}$$

K_i è una costante che deve essere $\neq 1$ e ha i che assume valore 0 e 1.
Sto differenziando la gravità di due tipi di errore con K_i, dato che la gravità dei due tipi di errore è diversa.

$$L(\theta,a_i) = \begin{cases} 0 & \text{se } \theta \in \Theta_i \\ K_i(\theta) & \text{se } \theta \in \Theta_j \end{cases}$$

6.2 Teoria delle decisioni statistiche

Le $a \in \mathcal{A}$ vengono scelte e ho un esperimento casuale ε e una $n-upla$ campionaria $\underline{x}$.
Ci deve essere un legame tra azioni e informazioni campionarie.

$$\textbf{(funzione di decisione) } \delta(\underline{x}) = a \textbf{ (stima)}$$

Voglio una regola che associa la $n-upla$ campionaria $(\underline{x})$ alla stima: $\rightarrow$ **Stimatori, Statistica test, Previsori**.

$$\delta(\underline{x}) = a \Rightarrow \Delta$$

con Δ che corrisponde alla classe di funzioni di decisione (serve solo in ambito classico) e contiene tutti i possibili stimatori.

6.2.1 Approccio teorico decisionale statistico classico

$\underline{x} \rightarrow \underline{X}$ (campionamento ripetuto).

$$L(\theta, a) = L(\theta, \delta(\underline{x}))$$

$$\text{rischio normale } R(\theta, \delta) = E^f L[(\theta, \delta(\underline{X}))]$$

$L(\theta, \delta(\underline{X}))$ è una variabile casuale e $\underline{X}$ è una funzione di verosimiglianza, legge congiunta.

Ho ora una terna (in ambito classico):

$$(\Theta,\ \Delta = \mathcal{A},\ R(\theta, \delta) = L(\theta, a))$$

$R(\theta, \delta)$ è la perdita quadratica, per la stima è MSE $\forall\theta$, spero che $\nexists$ uno stimatore ammissibile.

6.2.2 Approccio teorico decisionale statistico bayesiano

In ambito bayesiano abbiamo una quaterna:

$$(\Theta,\ \Delta = \mathcal{A},\ L(\theta, \delta(\underline{x})) \sim L(\theta, a),\ K)$$

con Θ supporto di θ, $\mathcal{A}$ contiene $a = \delta(\underline{x})$ che corrispondono a 2 numeri e K (valore atteso).

Perdita attesa finale

$$E^{\pi(\cdot|\underline{x})[L(\theta,\delta(\underline{x}))]} = \text{ PAF } \int_{\Theta} L(\theta, \delta(\underline{x})) \cdot \pi(\theta|\underline{x}) \, d\theta$$

altrimenti posso utilizzare la sommatoria ($\sum$).
So che θ è una variabile casuale poichè siamo in ambito bayesiano e $\delta(\underline{x})$ è un numero.

$$\rho(\pi(\cdot|\underline{x}), a) = \rho(\pi(\cdot|\underline{x}), \delta)$$

Inferenza bayesiana: cerca $\tilde{a}$ (oppure $\tilde{\delta}$) ottima:

$$\rho(\pi(\cdot|\underline{x}), \tilde{a}) \leq \rho(\pi(\cdot|\underline{x}), a) \ \forall a \in \mathcal{A}$$

$$\rho(\pi(\cdot|\underline{x}), \tilde{\delta}) \leq \rho(\pi(\cdot|\underline{x}), \delta) \ \forall \delta \in \Delta$$

Vediamo 2 modi di procedere in ambito bayesiano:

1) Analisi in forma estensiva

$$\tilde{\delta}_E \Rightarrow \tilde{\Delta}_E \subseteq \Delta \text{ se non è unica, metto tutto in una classe}$$

$\tilde{\delta}_E$ è ottenuto dalla procedura appena vista e non è unica.
$\underline{x} \Rightarrow \underline{X}$ voglio calcolare lo stimatore ottimo $\forall \ \underline{x} \in \mathcal{X}^{(n)}$.
(tipo di anaisi non necessaria in ambito bayesiano).

2) Analisi in forma normale

Abbiamo a disposizione π, L e $\underline{x} \Rightarrow \underline{X}$.

$$\text{rischio normale } R(\theta, \delta) : E^f[L(\theta, \delta(\underline{x}))]$$

faccio la media di $\underline{x}$ che poi verrà tolta e $\delta(\underline{x})$ è aleatoria.
Se non trovo una soluzione uniformemente migliore, vado su θ aleatorio e uso la prior $\pi(\theta)$. Effettuerò poi la media con approccio bayesiano.

$$E^{\pi(\theta)}[R(\theta, \delta)] = \text{ Rischio di Bayes} = r(\pi, \delta)$$

Voglio il δ che minimizza il rischio di Bayes.

$$\tilde{\delta}_N \Rightarrow \tilde{\Delta}_N$$

con $\tilde{\delta}_N$ che corrisponde a quello che minimizza e non è detto che sia unica.

Teorema:
Sotto condizioni di regolarità $\tilde{\Delta}_E$ e $\tilde{\Delta}_N$ sono uguali quasi certamente.

Dimostrazione:

$$r(\pi, \delta) = \int_\Theta R(\theta, \delta)\pi(\theta)\ d\theta$$

sostituisco poi $R(\theta, \delta)$ e ottengo:

$$\int_\Theta \int_{\mathcal{X}^{(n)}} L(\theta, \delta(\underline{x}))\ f(\underline{x}|\theta)\ d\underline{x}\ \pi(\theta)\ d\theta$$

$f(\underline{x}|\theta)$ dovrebbe essere $f(\underline{x}; \theta)$ poichè per questo passaggio ci troviamo nel caso classico.
Se posso, scambio gli integrali.
Per la condizione di regolarità ho:

$$\int_{\mathcal{X}^{(n)}} \int_\Theta L(\theta, \delta(\underline{x}))\ f(\underline{x}|\theta)\ \pi(\theta)\ d\theta\ d\underline{x}$$

con $f(\underline{x}|\theta)\ \pi(\theta)$ che corrisponde a $\pi(\theta|\underline{x}) \cdot m(\underline{x})$

$$\int_{\mathcal{X}^{(n)}} \int_\Theta L(\theta, \delta(\underline{x}))\ \pi(\theta|\underline{x})\ d\theta\ m(\underline{x})\ d\underline{x}$$

do come peso la posterior e poi integro

$$\int_\Theta L(\theta, \delta(\underline{x}))\ \pi(\theta|\underline{x})\ d\theta = E^{\pi(\cdot|\underline{x})}[L(\theta, \delta(\underline{x})] = \rho(\pi(\theta|\underline{x}), \delta(\underline{x})) = \text{ PAF}$$

$$r(\pi, \delta) = \int_{\mathcal{X}^{(n)}} \rho(\pi(\cdot|\underline{x}), \delta(\underline{x})\ m(\underline{x})\ d\underline{x}$$

Considero quindi $r(\pi, \delta) = \tilde{\delta}_N$ e $\rho(\pi(\cdot|\underline{x})) = \tilde{\delta}_E$ e ottengo

$$\tilde{\Delta}_E = \tilde{\Delta}_N$$

Minimizzando PAF

6.3 Stima puntuale

Vediamo come si fa inferenza bayesiana con approccio decisionale cercando di minimizzare la perdita attesa.
A "buon senso" per fare inferenza sulla posterior calcolata $\pi(\theta|\underline{x})$ posso prendere una qualunque sintesi. In realtà non tutte vanno bene, ma devo scegliere la sintesi migliore che mi dia stime ottime.
Quando parlo di stima bayesiana vuol dire che sto minimizzando un certo criterio, in particolare la perdita media ρ e la soluzione dipende dalla funzione di perdita che utilizzo:

$$\text{Ottimo } (min\ K = E = \text{ valore atteso} \Rightarrow \text{ quella che minimizza } \rho)$$

1) Perdita quadratica

$$L(\theta, a) = (\theta - a)^2 \rightarrow \rho(\pi(.|\underline{x}), a) = E^{\pi(.|\underline{x})}[(\theta - a)^2]$$

ora aggiungo e sottraggo $E^{\pi(.|\underline{x})}(\theta)$:

$$E^{\pi(.|\underline{x})}[\theta - E^{\pi(.|\underline{x})}(\theta)]^2 + E^{\pi(.|\underline{x})}[E^{\pi(.|\underline{x})}(\theta) - a]^2 + 2E^{\pi(.|\underline{x})}[\theta - E^{\pi(.|\underline{x})}(\theta)][E^{\pi(.|\underline{x})}(\theta) - a]$$

$$\longrightarrow \tilde{a} : \rho(E^{\pi(.|\underline{x})}, \tilde{a}) \leq \rho(E^{\pi(.|\underline{x})}, a) \forall a \in \mathcal{A}$$

Osservo che il terzo addendo è pari a 0: $2E^{\pi(.|\underline{x})}[\theta - E^{\pi(.|\underline{x})}(\theta)][E^{\pi(.|\underline{x})}(\theta) - a] = 0$, il primo non dipende da a mentre il secondo sì, ma poichè si tratta di un quadrato questo è sempre maggiore o al più uguale a 0, dunque l'azione ottima con perdita quadratica si avrà quando il secondo addendo è pari a 0, cioè:

$$E^{\pi(.|\underline{x})}[E^{\pi(.|\underline{x})}(\theta) - a]^2 = 0 \Leftrightarrow [E^{\pi(.|\underline{x})}(\theta) - a]^2 = 0$$
$$\Rightarrow \tilde{a} = E^{\pi(.|\underline{x})}(\theta)$$

L'azione ottima con perdita quadratica è la media a posteriori.

Nonostante la minimizzazione ho ancora

$$\rho(\pi(.|\underline{x}), \tilde{a}) = E^{\pi(.|\underline{x})}[(\theta - a)^2] = Var^{\pi(.|\underline{x})}(\theta)$$

questa è la **perdita minima inevitabile**.

2) Perdita quadratica ponderata

$$L(\theta, a) = \omega(\theta)(\theta - a)^2$$

Cambiano le soluzioni in base alla funzione di peso.

$$\longrightarrow \rho(\pi(.|\underline{x}), a) = E^{\pi(.|\underline{x})}[\omega(\theta)(\theta - a)^2] = \int_\Theta \omega(\theta)(\theta - a)^2 \pi(\theta|\underline{x}) d\theta$$

L'integrale è l'oggetto che vogliamo minimizzare.
Per trovare $\tilde{a}$ calcolo

$$\frac{\delta\rho}{\delta a} \rightarrow \tilde{a} = \frac{E^{\pi(.|\underline{x})}[\omega(\theta)\cdot\theta]}{E^{\pi(.|\underline{x})}[\omega(\theta)]}$$

Osservazione: se $\omega(\theta) = \omega \Rightarrow \tilde{a} = E^{\pi(.|\underline{x})}(\theta)$ ottengo il risultato visto per la perdita quadratica cioè la **media a posteriori**, invece quando $\omega(\theta) = \frac{1}{\theta} \Rightarrow \tilde{a} = \frac{1}{E^{\pi(.|\underline{x})}(\frac{1}{\theta})}$ ottengo la **media armonica**.

3) Perdita lineare asimmetrica

$$L(\theta, a) = \begin{cases} k_0|\theta - a| & \text{se } \theta \geq a \\ k_1|a - \theta| & \text{se } \theta < a \end{cases}$$

dove k_0 e k_1 sono costanti positive che mi permettono di differenziare tra sovrastima e sottostima, decido io la gravità di questi errori in base al peso che gli do.
Se $k_0 = k_1$ ho i **risultati validi per la perdita assoluta simmetrica** ($L(\theta, a) = |\theta - a|$).

$$\rho(\pi(.|\underline{x}), a) = E^{\pi(.|\underline{x})}[L(\theta, a)] = \int_{-\infty}^{a} k_1|a-\theta|\pi(\theta|\underline{x})d\theta + \int_{a}^{+\infty} k_0|\theta-a|\pi(\theta|\underline{x})d\theta$$

Ora devo minimizzare rispetto ad a per trovare $\tilde{a}$ (il problema è che a compare anche nell'integrale).

In generale:

$$\frac{\delta}{\delta t}\int_{b(t)}^{c(t)} g(t;s)ds = \int_{b(t)}^{c(t)} \frac{\delta}{\delta t} g(t;s)ds + c'(t)g(c(t),s) - b'(t)g(b(t),s)$$

Obbiettivo:

$$\frac{\delta\rho}{\delta a} = 0 \rightarrow \frac{\delta\rho}{\delta a}\int_{-\infty}^{a} k_1\pi(\theta|\underline{x})d\theta + 1\cdot(a-a) - 0 - \int_{a}^{+\infty} k_0\pi(\theta|\underline{x})d\theta + 0 - 0 = 0$$

Ora devo risolvere l'equazione,aggiungo e tolgo: $\pm\int_{-\infty}^{a} k_0\pi(\theta|\underline{x})d\theta$

$$\rightarrow -k_0\int_{-\infty}^{+\infty}\pi(\theta|\underline{x})d\theta + (k_0+k_1)\int_{-\infty}^{a}\pi(\theta|\underline{x})d\theta = 0$$

Sapendo che ho la legge di distribuzione: $\int_{-\infty}^{+\infty}\pi(\theta|\underline{x})d\theta = 1$

$$\text{e: } \int_{-\infty}^{a} \pi(\theta|\underline{x})d\theta = \frac{k_0}{k_0+k_1}$$

In pratica devo integrare la posterior fino ad $\tilde{a}$ e osservo che $\frac{k_0}{k_0+k_1} \in (0,1) \Rightarrow$ **$\tilde{\mathbf{a}}$**

$$\textbf{è il quantile di ordine } \frac{\mathbf{k_0}}{\mathbf{k_0+k_1}}$$

Osservazione:
Se $k_0 = k_1$ ho il caso simmetrico (perdita assoluta)
$\rightarrow \tilde{a} = \theta_{0.5}$ (mediana).
Bisogna stare attenti a scegliere k_0 e k_1, è meglio optare per la scelta simmetrica.

4) Perdita 0/1

$$L(\theta,a) = \begin{cases} 0 & |\theta-a| \leq \varepsilon \quad \varepsilon > 0 \\ 1 & |\theta-a| > \varepsilon \end{cases}$$

$$\rho(\pi(.|\underline{x}),a) = E^{\pi(.|\underline{x})}[L(\theta,a)] = 0 \cdot P^{\pi(.|\underline{x})}(|\theta-a| \leq \epsilon) + 1 \cdot P^{\pi(.|\underline{x})}(|\theta-a| > \epsilon)$$
$$= 1 - P^{\pi(.|\underline{x})}(|\theta-a| \leq \epsilon)$$

Allora voglio trovare:

$$\tilde{a} : min(\rho) \rightarrow max[P^{\pi(.|\underline{x})}(|\theta-a| \leq \epsilon)]$$

Questa condizione si verifica quando $(a-\epsilon, a+\epsilon)$ ha probabilità massima

$$\Rightarrow \tilde{a} = \textbf{moda a posteriori}$$

Riepilogo:

$\mathbf{L}(\theta, \mathbf{a})$	$\tilde{\mathbf{a}}$
Perdita quadratica	**Media a posteriori**
Perdita assoluta	**Mediana a posteriori**
Perdita lineare asimmetrica	**Quantile** $\frac{\mathbf{k_0}}{\mathbf{k_0+k_1}}$
Perdita 0/1	**Moda a posteriori**

Rispetto alla funzione di perdita si vuole robustezza, cioè quando questa viene cambiata la stima varia di "poco".

Osservazione:
Se $\pi(\theta|\underline{x})$ è simmetrica e unimodale la media e la mediana coincidono
$\Rightarrow$ ho robustezza rispetto alla funzione di perdita.

Osservazione:
In base alla scelta della prior ($\pi(\theta)$) posso ottenere la robustezza rispetto alla scelta di quest'ultima.

Domanda:
Supponi d scegliere la prior di Laplace mentre la funzione di perdita è 0/1, facendo la stima bayesiana di $\tilde{a}$ si ottiene un risultato noto?

$$\tilde{a} = \theta : max[\pi(\theta|\underline{x})] \Longrightarrow \tilde{a} = \text{Stima di Massima Verosimiglianza } \textbf{(SVM)}$$

6.3.1 Stima puntuale per parametri multipli

Ora non ho più il parametro θ ma $\underline{\theta} = (\theta_1, ..., \theta_k)$.

Con $\underline{a}$ vettore ho che

$$L(\underline{\theta}, \underline{a}) = (\underline{\theta} - \underline{a})^T Q(\underline{\theta} - \underline{a})$$

è una generalizzazione della perdita quadratica ponderata, dove Q è una matrice $k \cdot k$ simmetrica e definita positiva (rappresenta i pesi per i prodotti degli errori di stima).

$$\rho(\pi(.|\underline{x}), \underline{a}) = E^{\pi(.|\underline{x})}[(\underline{\theta} - \underline{a})' Q(\underline{\theta} - \underline{a})]$$

aggiungo e tolgo $\pm E^{\pi(.|\underline{x})}(\theta)$ dall'equazione e ottengo:

$$E^{\pi(.|\underline{x})}\{(\theta - E^{\pi(.|\underline{x})}(\theta) + E^{\pi(.|\underline{x})}(\theta) - \underline{a})' Q(\theta - E^{\pi(.|\underline{x})}(\theta) + E^{\pi(.|\underline{x})}(\theta) - \underline{a})\} =$$

$$E^{\pi(.|\underline{x})}\{[\theta - E^{\pi(.|\underline{x})}(\theta)]' Q[\theta - E^{\pi(.|\underline{x})}(\theta)] + [E^{\pi(.|\underline{x})}(\theta) - \underline{a}]' Q[E^{\pi(.|\underline{x})}(\theta) - \underline{a}] + 2 \cdot 0\} =$$

Il primo addendo non dipende da $\underline{a}$ mentre nel secondo addendo compare ed è sempre positivo tranne quando $\underline{a} = E^{\pi(.|\underline{x})}(\theta)$

$$\Longrightarrow \tilde{a} = \text{vettore delle medie della posterior } \forall\ Q$$

Si nota che la soluzione non dipende da Q ma questa influisce su $\rho[pi(.|\underline{x}), \underline{a}]$, cioè quanto perdo; in particolare quello che perdo è la traccia di $\Sigma_{\theta|\underline{x}}$, cioè la somma delle varianze a posteriori.

6.3.2 Stima puntuale per trasformate del parametro

$$\underline{\theta} = \begin{cases} \omega = \tau(\underline{\theta}) & \text{oggetto di inferenza (parametro di interesse)} \\ \lambda = h(\underline{\theta}) & \text{non oggetto di inferenza (parametro di disturbo)} \end{cases}$$

Ad esempio nello studio di una Normale ho $\underline{\theta} = (\mu, \sigma^2)$ e il parametro su cui voglio fare inferenza (parametro di interesse) è $\omega = \mu$, $h(\underline{\theta})$ lo posso scegliere arbitrariamente, è conveniente fare in modo che la trasformata sia biunivoca.

Esempio:
Con $x \sim Poisson(\theta_1)$ e $y \sim Poisson(\theta_2)$ ho quindi il vettore di parametri pari a $\underline{\theta} = (\theta_1, \theta_2)$.
$\Rightarrow \omega = \tau(\underline{\theta}) = \frac{\theta_1}{\theta_2}$ non è una trasformata biunivoca di $\underline{\theta}$.

Si hanno problemi quando ω è una trasformata non biunivoca di $\underline{\theta}$ per $\lambda = h(\underline{\theta})$ scelgo una trasformata complementare.

$\rightarrow f(\underline{x}|\underline{\theta})$:
• $\underline{\theta} = (\omega, \lambda)$ rispettivamente parametro di interesse e parametro di disturbo.
• riparametrizzo $f(\underline{x}|(\omega, \lambda))$
• calcolo la funzione di verosimiglianza che dipende solo da ω (pseudo verosimiglianza).

Mi serve un passaggio dalla verosimiglianza effettiva alla pseudo verosimiglianza (in pratica devo togliere λ) che ha l'utilizzo tipico della verosimiglianza.

Frequentista:
partendo dalla funzione di verosimiglianza $f(x; \omega, \lambda)$ voglio arrivare alla verosimiglianza profilo $\mathbf{f_{PV}}(\underline{\mathbf{x}}; \omega)$. Allora devo calcolare

$$sup_\lambda f(x; \omega, \lambda) = \hat{\lambda}(\omega) \rightarrow f(\underline{x}; \omega, \hat{\lambda}(\omega))$$

In ambito bayesiano la pseudo verosimiglianza si chiama verosimiglianza integrata. Vediamo come si calcola a partire da $(\omega, \lambda) \rightarrow \pi(\omega, \lambda|\underline{x})$.
Per eliminare λ devo marginalizzare rispetto a quest'ultima:

$$\pi(\omega|\underline{x}) = \int_\Lambda \pi(\omega, \lambda|\underline{x}) d\lambda \propto \int_\Lambda \pi(\omega, \lambda) f(\underline{x}|\lambda, \omega) d\lambda$$

A questo punto se la prior è propria posso scrivere:

$$\pi(\omega)\int_{\Lambda}\pi(\lambda|\omega)f(\underline{x}|\omega,\lambda)d\lambda$$

Questa funzione non dipende da λ ed ha il ruolo della verosimiglianza $\Rightarrow$ è la **verosimiglianza integrata o pseudo verosimiglianza**.

ESERCIZIO (Poisson)
Sia $X \sim Po(\theta_1)$ e $Y \sim Po(\theta_2)$, il parametro di interesse è $\omega = \frac{\theta_1}{\theta_2}$.
Suppongo di avere una prova in X e una in Y.

$$\begin{cases}\omega = \frac{\theta_1}{\theta_2} & \text{parametro di interesse}\\ \lambda = \theta_2 & \text{parametro di disturbo (è una trasformazione complementare)}\end{cases}$$

$$f(x,y|\underline{\theta}) = \frac{e^{-\theta_1}\theta_1^x}{x!}\frac{e^{-\theta_2}\theta_2^y}{y!}$$

Ora riparametrizzo:

$$\begin{cases}\theta_1 = \omega\cdot\lambda\\ \theta_2 = \lambda\end{cases} \longrightarrow f(x,y|\omega,\lambda) = \frac{e^{-\omega\lambda-\lambda}(\omega\lambda)^x\lambda^y}{x!\cdot y!}$$

$$\pi(\omega) = \frac{1}{\sqrt{\omega(1+\omega)}}$$

$$\pi(\lambda|\omega) = \frac{1}{\sqrt{\lambda}}$$

Allora la verosimiglianza integrata è:

$$\int_{\Lambda}\pi(\lambda|\omega)f(x,y|\omega,\lambda)d\lambda = \int_{\Lambda}\frac{1}{\sqrt{\lambda}}e^{-\lambda(1+\omega)}\lambda^{x+y}\omega^x d\lambda = \omega^x\int_{\Lambda}\lambda^{x+y-\frac{1}{2}}e^{-\lambda(1+\omega)}d\lambda$$

Nell'integrale riconosco il nucleo di una $Ga(x+y+\frac{1}{2}, 1+\omega)$, allora posso riscrivere il tutto come

$$\frac{\omega^x\Gamma(x+y+\frac{1}{2})}{(1+\omega)^{x+y+\frac{1}{2}}} \propto \frac{\omega^x}{(1+\omega)^{x+y+\frac{1}{2}}}$$

questa è la verosimiglianza integrata.

6.4 Stima intervallare

Scelgo un insieme $S \subseteq \Theta$.

[$HPD : S_h = \{\theta : \pi(\theta|\underline{x}) \geq h\}$]

Ora vediamo l'**HPD esteso:**

$$\textbf{EHPD} : S_h^{'} = \{\theta : \pi(\theta|\underline{x}) > h\} \in S_h^{'}$$

$$\textbf{EHPD} : S_h^{'} = \{\theta : \pi(\theta|\underline{x}) < h\} \notin S_h^{'}$$

$$\textbf{EHPD} : S_h^{'} = \{\theta : \pi(\theta|\underline{x}) = h\}$$

Non vengono fatte richieste sugli estremi.
$L(\theta, S) \rightarrow$ se ho intervalli grossi ci perdo, lo stesso se S non contiene il vero valore di θ.
$L(\theta, S) = F(mis(S)) - I(\theta) \longrightarrow$ F è una trasformata monotona lineare della misura di S e tolgo l'indicatore del vero valore del parametro.
Ad esempio $L(\theta, S) = b(mis(S)) - I(\theta)$ con $b > 0$.

Osservazione: se L è monotona e θ è discreta allora tutte le $\tilde{S}$ sono EHPD.
Osservazione: se L è monotona e θ è continua allora $\exists$ $\tilde{S}$ che è EHPD.
A "buon senso" si sintetizzano le scelte bayesiane ottime con gli HPD.

7 Verifica d'ipotesi

7.1 Senza approccio decisionale

$$H_0 : \theta \in \Theta_0 \qquad H_1 : \theta \in \Theta_1$$

Rifiuto H_0 se $P(\theta \in \Theta_1|\underline{x}) > P(\theta \in \Theta_0|\underline{x})$.

7.2 Con approccio decisionale

$$\mathcal{A} = \{a_0, a_1\}$$

Con a_0 se accetto H_0 e a_1 se rifiuto H_0.

Prinicpali funzioni di perdita: (perdita 0/1 e varianti)

$$L(\theta, a_i) = \begin{cases} 0 & \text{se } \theta \in \Theta_i \\ 1 & \text{se } \theta \in \Theta_j \end{cases}$$

Con $i = 0, 1$ e $j = 0, 1$.

$$L(\theta, a_i) = \begin{cases} 0 & \text{se } \theta \in \Theta_i \\ K_i & \text{se } \theta \in \Theta_j \end{cases}$$

K_i è una costante che deve essere $\neq 1$ e ha i che assume valore 0 e 1.
Sto differenziando la gravità di due tipi di errore con K_i, dato che la gravità dei due tipi di errore è diversa.

$$L(\theta, a_i) = \begin{cases} 0 & \text{se } \theta \in \Theta_i \\ K_i(\theta) & \text{se } \theta \in \Theta_j \end{cases}$$

Azione ottima ($\tilde{a}$) con perdita 0/1:

$$\rho(\pi(.|\underline{x}), a_0) = E^{\pi(.|\underline{x})}[L(\theta, a_0)]$$

$$= 0 \cdot P(\theta \in \Theta_0|\underline{x}) + 1 \cdot P(\theta \in \Theta_1|\underline{x}) = P(\theta \in \Theta_1|\underline{x})$$

con procedimento analogo:

$$\rho(\pi(.|\underline{x}), a_1) = E^{\pi(.|\underline{x})}[L(\theta, a_1)]$$

$$= 1 \cdot P(\theta \in \Theta_0|\underline{x}) + 0 \cdot P(\theta \in \Theta_1|\underline{x}) = P(\theta \in \Theta_0|\underline{x})$$

rifiuto H_0 se

$$\rho(\pi(.|\underline{x}), a_0) > \rho(\pi(.|\underline{x}), a_1)$$

minimizzo ρ, quindi se

$$P(\theta \in \Theta_0|\underline{x}) < P(\theta \in \Theta_1|\underline{x})$$

$\Rightarrow$ **Stessa conclusione dell'approccio non decisionale (a buon senso).**

Con passaggi analoghi si ha che se L è di tipo $0/K_i$ (differenzio la perdita in base al tipo di errore $\rightarrow$ cosa più realistica) allora rifiuto H_0 se

$$\frac{P(\theta \in \Theta_0|\underline{x})}{P(\theta \in \Theta_1|\underline{x})} < \frac{K_0}{K_1}$$

Osservazione:
Se Θ_0 e Θ_1 sono una partizione di Θ allora

$$P(\theta \in \Theta_0|\underline{x}) = 1 - P(\theta \in \Theta_1|\underline{x})$$

Quindi, sostituendo, si ha che rifiuto se

$$P(\theta \in \Theta_1|\underline{x}) > \frac{K_1}{K_0 + K_1}$$

Al solito si pone in via naturale la domanda:
Quando il test bayesiano (ottimo) coincide con il test più potente (o uniformemente più potente) della statistica classica?

7.2.1 Caso 1

Le ipotesi sono semplici:

$$H_0 : \theta = \theta_0 \qquad H_1 : \theta = \theta_1$$

non ci sono parametri di disturbo.
Indichiamo le prior con:

$$\pi(\theta_0) = P \qquad \pi(\theta_1) = 1 - P \qquad (0 < P < 1)$$

allora

$$P(\theta \in \Theta_1|\underline{x}) = \frac{(1-P) \cdot f(\underline{x}|\theta_1)}{Pf(\underline{x}|\theta_0) + (1-P)f(\underline{x}|\theta_1)}$$

il rifiuto si ha per

$$(1-P)f(\underline{x}|\theta_1) > \frac{K_1}{K_0 + K_1}[Pf(\underline{x}|\theta_0) + (1-P)f(\underline{x}|\theta_1)]$$

e con qualche passaggio algebrico:

$$\frac{f(\underline{x}|\theta_1)}{f(\underline{x}|\theta_0)} > \frac{K_1}{K_0} \cdot \frac{P}{1-P}$$

dove $\frac{f(\underline{x}|\theta_1)}{f(\underline{x}|\theta_0)}$ è il rapporto tra la verosimiglianza sotto H_1 e quella sotto H_0.

Il test più potente (Neyman-Pearson) ha la zona di rifiuto

$$\frac{f(\underline{x};\theta_1)}{f(\underline{x};\theta_0)} > K_\alpha$$

dove α è la probabilità dell'errore di prima specie $\Rightarrow$ La struttura è la stessa (rifiuto se il rapporto tra le verosimiglianze è "troppo grande").

Classica $\rightarrow$ la soglia dipende da α.
Bayesiana $\rightarrow$ la soglia dipende dalla prior (P) e dalla funzione di perdita (K_0, K_1).

7.2.2 Caso 2

Le ipotesi sono composte unidirezionali: non posso dimostrare nulla in generale, ma nella maggior parte dei casi i test bayesiani hanno una struttura "classica".

Esercizio 1
n prove indipendenti in $X \sim N(\theta, \sigma^2 = \text{ nota})$

$$H_0 : \theta \leq \theta_0 \qquad H_1 : \theta > \theta_0$$

[il test uniformemente più potente ha quale zona di rifiuto: $\{\underline{x} : \overline{x} \geq K_\alpha\}$] con $\overline{x}$ media campionaria.
Si trovi il test ottimo bayesiano con $\pi(\theta) \propto 1$ e la funzione di perdita $0/K_i$.

Soluzione esercizio 1
Posterior $\Rightarrow N(\overline{x}, \frac{\sigma^2}{n})$

$$\rho(\pi(.|\underline{x}), a_0) = K_0 \cdot P(\theta \in \Theta_1|\underline{x}) = K_0 \int_{\theta_0}^{\infty} \frac{1}{\sqrt{2\pi}} \frac{\sqrt{n}}{\sigma} exp\{-\frac{m}{2\sigma^2}(\theta - \overline{x})^2\}\, d\theta$$

$$= K_0 \int_{\frac{(\theta_0-\overline{x})\sqrt{n}}{\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} exp\{-\frac{z^2}{2}\}\, dz = K_0\{1 - \Phi_z(\frac{\theta_0 - \overline{x}}{\sigma}\sqrt{n})\}$$

con Φ_z che è la funzione di ripartizione di una $N(0,1)$.

Con analoghi passaggi:

$$\rho(\pi(.|\underline{x}), a_1) = K_1 \cdot P(\theta \in \Theta_0|\underline{x}) = K_1 \int_{-\infty}^{\theta_0} \frac{1}{\sqrt{2\pi}} \frac{\sqrt{n}}{\sigma} exp\{-\frac{m}{2\sigma^2}(\theta - \overline{x})^2\}\ d\theta$$

$$= K_1 \int_{-\infty}^{\frac{(\theta_0 - \overline{x})\sqrt{n}}{\sigma}} \frac{1}{\sqrt{2\pi}} exp\{-\frac{z^2}{2}\}\ dz = K_1\{\Phi_z(\frac{\theta_0 - \overline{x}}{\sigma}\sqrt{n})\}$$

Quindi si rifiuta se

$$K_1\{\Phi_z(\frac{\theta_0 - \overline{x}}{\sigma}\sqrt{n}) < K_0\{\Phi_z(1 - \frac{\theta_0 - \overline{x}}{\sigma}\sqrt{n})\}\}$$

cioè se:

$$\overline{x} > \theta_0 - \frac{\sigma}{\sqrt{n}}\Phi^{-1}(\frac{K_0}{K_0 + K_1})$$

La media campionaria è "troppo grande" come nel test classico.

Esercizio 2
$X \sim N(m = \text{ noto}, \sigma^2 = \frac{1}{\theta})$
$\pi(\theta) \propto \frac{1}{\theta}$

$$H_0 : \theta \leq \theta_0 \qquad H_1 : \theta > \theta_1$$

La funzione di perdita è $0/K_i$.

Trovare il test ottimo bayesiano e confrontarlo con quello classico che ha zona critica

$$\sum_{i=1}^{n}(x_i - m)^2 < K_\alpha$$

Soluzione esercizio 2
La posterior è una

$$Gamma\left(\frac{n}{2}, \frac{\sum_{i=1}^{n}(x_i - m)^2}{2}\right)$$

$$\rho(\pi(.|\underline{x}), a_0) = K_0\{1 - \Psi(\sum_{i=1}^{n}(x_i - m)^2\theta_0)\}$$

con Ψ funzione di ripartizione di una χ^2_n

$$\rho(\pi(.|\underline{x}), a_1) = K_1\{\Psi(\sum_{i=1}^{n}(x_i - m)^2\theta_0)\}$$

Si rifiuta se

$$\sum_{i=1}^{n}(x_i - m)^2\theta_0) < \frac{1}{\theta_0}\Psi^{-1}(\frac{K_0}{K_0 + K_1})$$

7.2.3 Caso 3

Ipotesi alternativa bidirezionale (H_0 "precisa").
I risultati bayesiani sono in genere diversi da quelli classici.
In realtà non ha senso confrontare

$$H_0 : \theta = \theta_0 \qquad H_1 : \theta \neq \theta_0$$

1 punto per l'ipotesi nulla e tutti gli altri per l'ipotesi alternativa.
L'ipotesi nulla è da interpretare come un intorno di θ_0 di cui H_0 è un'approssimazione.

La prior per θ dovrà essere una legge che mette su θ_0 una probabilità $P > 0$ (altrimenti se $P = 0 \Rightarrow \theta_0$ avrà probabilità nulla anche a posteriori!).
Quindi se Θ prevede un'infinità non numerabile di valori la prior sarà:

$$P(\theta \in \Theta_0) = P$$

$$P(\theta \in \Theta_1) = 1 - P$$

e $(1 - P)$ si diffonde su Θ_1 secondo una qualche densità $g(\theta)$.
Quindi:

$$P(\theta \in \Theta_0 | \underline{x}) = \frac{P \cdot f(\underline{x}|\theta_0)}{P \cdot f(\underline{x}|\theta_0) + (1 - P) \int_{\Theta_1} f(\underline{x}|\theta) g(\theta)\ d\theta}$$

$\int_{\Theta_1} f(\underline{x}|\theta) g(\theta)\ d\theta$ ha la stessa struttura della verosimiglianza integrata.

N.B. con gli usuali passaggi (calcolo ρ e la minimizzo) si ottiene il rifiuto di H_0 se

$$\frac{\int_{\Theta_1} f(\underline{x}|\theta) g(\theta)\ d\theta}{f(\underline{x}|\theta)} > \frac{K_1}{K_0} \frac{P}{1 - P}$$

$\Rightarrow$ stessa struttura vista per ipotesi semplici, ma $f(\underline{x}|\theta_1)$ è sostituito da $\int_{\Theta_1} f(\underline{x}|\theta) g(\theta)\ d\theta$.

7.3 Il fattore di Bayes

L'ODDS a favore di un'ipotesi H_i è:

$$O(H_i) = \frac{P(H_i)}{1 - P(H_i)} = \frac{P(H_i)}{P(\overline{H}_i)}$$

Se si è verificato un evento (nel nostro caso la n-upla è risultata $\underline{x}$) allora si hanno l'ODDS a priori e quello a posteriori:

$$O(H_i|\underline{x}) = \frac{P(H_i|\underline{x})}{P(\overline{H}_i|\underline{x})}$$

Il fattore di Bayes a favore di un'ipotesi (ad esempio H_0) in presenza del risultato sperimentale $\underline{x}$, è:

$$B(H_0|\underline{x}) = \frac{O(H_0|\underline{x})}{O(H_0)}$$

Quindi

$$O(H_0) \cdot B(H_0|\underline{x}) = O(H_0|\underline{x})$$

con $B(H_0|\underline{x})$ fattore per il quale moltiplicare l'ODDS a priori per pervenire a quello a posteriori.

Risulta che:

$$B(H_0|\underline{x}) = \frac{O(H_0|\underline{x})}{O(H_0)} = \frac{P(H_0|\underline{x})}{P(H_1|\underline{x})} \cdot \frac{P(H_1)}{P(H_0)}$$

$$= \frac{\int_{\Theta_0} f(\underline{x}|\theta_0)\pi(\theta)\ d\theta}{\int_{\Theta_1} f(\underline{x}|\theta_1)\pi(\theta)\ d\theta} \cdot \frac{\int_{\Theta_1} \pi(\theta)\ d\theta}{\int_{\Theta_0} \pi(\theta)\ d\theta}$$

Se si indica con

$$g_i(\theta) = \frac{\pi(\theta)\ I_{\Theta_i}(\theta)}{\int_{\Theta_i} \pi(\theta)\ d\theta}$$

La densità di θ condizionata a $\theta \in \Theta_i$, allora il fattore di bayes diventa:

$$B(H_0|\underline{x}) = \frac{\int_{\Theta_0} f(\underline{x}|\theta_0) g_0(\theta)\ d\theta}{\int_{\Theta_1} f(\underline{x}|\theta_1) g_1(\theta)\ d\theta}$$

è un rapporto tra 2 verosimiglianze integrate.

Osservazioni conclusive:
1) Se H_0 e H_1 sono semplici il $B(H_0|\underline{x})$ si riduce al rapporto tra le verosimiglianze sotto H_0 e sotto H_1 (concetto che viene interpretato da tutti come

supporto all'ipotesi fornita dai dati).
2) Dato che il fattore di Bayes fornisce una misura dell'evidenza a favore di H_i fornita da $\underline{x}$, spesso la verifica di ipotesi in ambito bayesiano viene condotta tramite il fattore di Bayes.
Ad esempio: Rifiuto H_0 se

$$K_1 P(H_0|\underline{x}) < K_0 P(H_1|\underline{x})$$

con $P(H_0|\underline{x}) \equiv P(\theta \in \Theta_0|\underline{x})$ e $P(H_1|\underline{x}) \equiv P(\theta \in \Theta_1|\underline{x})$.

Ovvero

$$O(H_0|\underline{x}) < \frac{K_0}{K_1}$$

Ovvero se

$$B(H_0|\underline{x})O(H_0) < \frac{K_0}{K_1}$$

Ovvero se

$$B(H_0|\underline{x}) < \frac{K_0}{K_1}\frac{P(H_1)}{P(H_0)}$$

Rifiuto se il fattore di Bayes a favore dell'ipotesi nulla è "troppo piccolo".