# Statistica Bayesiana

*Vers.* 1.1.1

Gianluca Mastrantonio

email: gianluca.mastrantonio@polito.it

# Il teorema di Bayes

# Il teorema di Bayes

La statistica bayesiana si fonda sul teorema di bayes, che dice che data due variabili aleatorie (anche vettoriali) $\mathbf{X}$ e $\mathbf{Y}$, allora

$$f(\mathbf{x}|\mathbf{y}) = \frac{f(\mathbf{y},\mathbf{x})}{f(\mathbf{y})} = \frac{f(\mathbf{y}|\mathbf{x})f(\mathbf{x})}{f(\mathbf{y})}$$

dove $f(\mathbf{y}) = \int f(\mathbf{y},\mathbf{x})d\lambda(x)$. Il teorema di Bayes permette di passare dalla condizionata di $\mathbf{y}|\mathbf{x}$ a quella di $\mathbf{x}|\mathbf{y}$.

Un altro modo di vedere il teorema di Bayes è di tipo "iterativo"

- ho una distribuzione a-priori su $\mathbf{x}$, $f(\mathbf{x})$
- Osservo una nuova variabile $\mathbf{y}$, che dipende da $\mathbf{x}$, tramite $f(\mathbf{y}|\mathbf{x})$
- allora l'informazione che ho su $\mathbf{x}$, dopo aver osservato $\mathbf{y}$, cambia in $f(\mathbf{x}|\mathbf{y})$.

# Il teorema di Bayes

Facciamo un esempio

**Binomiali**

Ipotizziamo di avere un dado a 6 facce e di avere il dubbio che sia truccato e tenda a far uscire più spesso numeri pari che dispari. Facciamo un esperimento, lanciamo il dado n volte, e troviamo che per $y$ volte è uscito un pari e $n - y$ un dispari. Abbiamo abbastanza evidenze che sia truccato?

**Soluzione:**

Usiamo il teorema di bayes e assumiamo che $\theta$ sia il parametro di interesse e indichi la probabilità che esca un numero pari, mentre $\mathbf{Y}^* = (Y_1^*, \dots Y_n^*)'$ è il vettore con i risultati dell'esperimento e $Y = \sum Y_i^*$ è la loro somma. Siamo quindi interessati a

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)}$$

# Il teorema di Bayes

Sappiamo che

$$Y|\theta \sim Bin(n, \theta)$$

e per la distribuzione di $\theta$ assumiamo una beta

$$\theta \sim Beta(a, b)$$

Non ci resta che calcolare $f(\theta|y)$. Partiamo dal numeratore che è uguale a

$$f(y|\theta)f(\theta) = \left( \begin{array}{c} n \\ y \end{array} \right) \theta^y (1-\theta)^{n-y} \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} = \left( \begin{array}{c} n \\ y \end{array} \right) \frac{\theta^{y+a-1}(1-\theta)^{n+b-y-1}}{B(a,b)}$$

dove

$$B(a,b) = \int_0^1 u^{a-1}(1-u)^{b-1} du$$

In questo caso il denominatore è l'integrale del numeratore e quindi

$$f(y) = \int_0^1 \left( \begin{array}{c} n \\ y \end{array} \right) \frac{u^{y+a-1}(1-u)^{n+b-y-1}}{B(a,b)} du = \frac{\left( \begin{array}{c} n \\ y \end{array} \right)}{B(a,b)} \int_0^1 u^{y+a-1}(1-u)^{n+b-y-1} du$$

# Il teorema di Bayes

Mettendo tutto insieme abbiamo che

$$f(\theta|y) = \frac{\theta^{y+a-1}(1-\theta)^{n+b-y-1}}{\int_0^1 u^{y+a-1}(1-u)^{n+b-y-1}du} = \frac{\theta^{y+a-1}(1-\theta)^{n+b-y-1}}{B(y+a, n+b-y)}$$

che è la densità di una $B(y+a, n+b-y)$. □

# Il teorema di Bayes

In statistica Bayesiana, la distribuzione $f(\boldsymbol{\theta}|\mathbf{y})$ è chiamata distribuzione a posteriori di $\boldsymbol{\theta}$, e si compone di tre elementi

- $f(\mathbf{y}|\boldsymbol{\theta})$: la congiunta della osservazioni, che è possibile vedere anche come la verosimiglianza;
- $f(\boldsymbol{\theta})$: la distribuzione a priori. Questa distribuzione riflette ciò che sappiamo dei parametri prima di osservare il campione $\mathbf{y}$, i.e. non dipende da $\mathbf{y}$;
- $f(\mathbf{y})$: costante di normalizzazione. In genere "poco importante" visto che non dipende da $\boldsymbol{\theta}$.

La scelta delle a-priori è molto importante e bisogna stare attenti. In generale, se possibile, si preferisce utilizzare distribuzioni che sono costanti o molto piatte, e.g.,

- $U(a, b)$ se la variabile è definita su $(a, b)$,
- oppure $N(0, 100000)$ se è definita su $\mathbb{R}$.

# Il teorema di Bayes

In questo caso è come se stessimo dicendo che non sappiamo, a-priori, che valore può assumere la variabile/parametro, e lasciamo decidere ai dati la a posteriori. Queste prior si chiamano **non-informative** o **debolmente informative** (hanno una varianza elevata e la densità/probabilità è approssimativamente costante).
Se usiamo prior che mettono molta probabilità su particolari valori

- $Beta(10, 10)$ se la variabile è definita su $(0, 1)$,
- oppure $N(5, 0.1)$ se è definita su $\mathbb{R}$.

allora stiamo mettendo molta informazione a priori, e la a posteriori dipenderà molto dalla a priori. Queste si chiamano prior **informative** (hanno una varianza bassa e ci sono punti con densità/probabilità elevata).

# Il teorema di Bayes

Per esempio, la media di una $Beta(a, b)$ è $\frac{a}{a+b}$ e la varianza è $\frac{ab}{(a+b)^2(a+b+1)^2}$. La Beta non informativa ha parametri $a = b = 1$ (è uniforme), mentre, un esempio di Beta informativa è $a = b = 1000$. Notate che entrambe hanno la stessa media.
La media della a posteriori è

$$\frac{y + a}{a + n + b}$$

che sarà simile come valore a $\frac{y}{n}$ con la prior non informativa (quindi i dati decidono la a posteriori), mentre sarà simile a $\frac{a}{a+b}$ con la prior informativa (la a posteriori è simile alla prior).

# Il teorema di Bayes

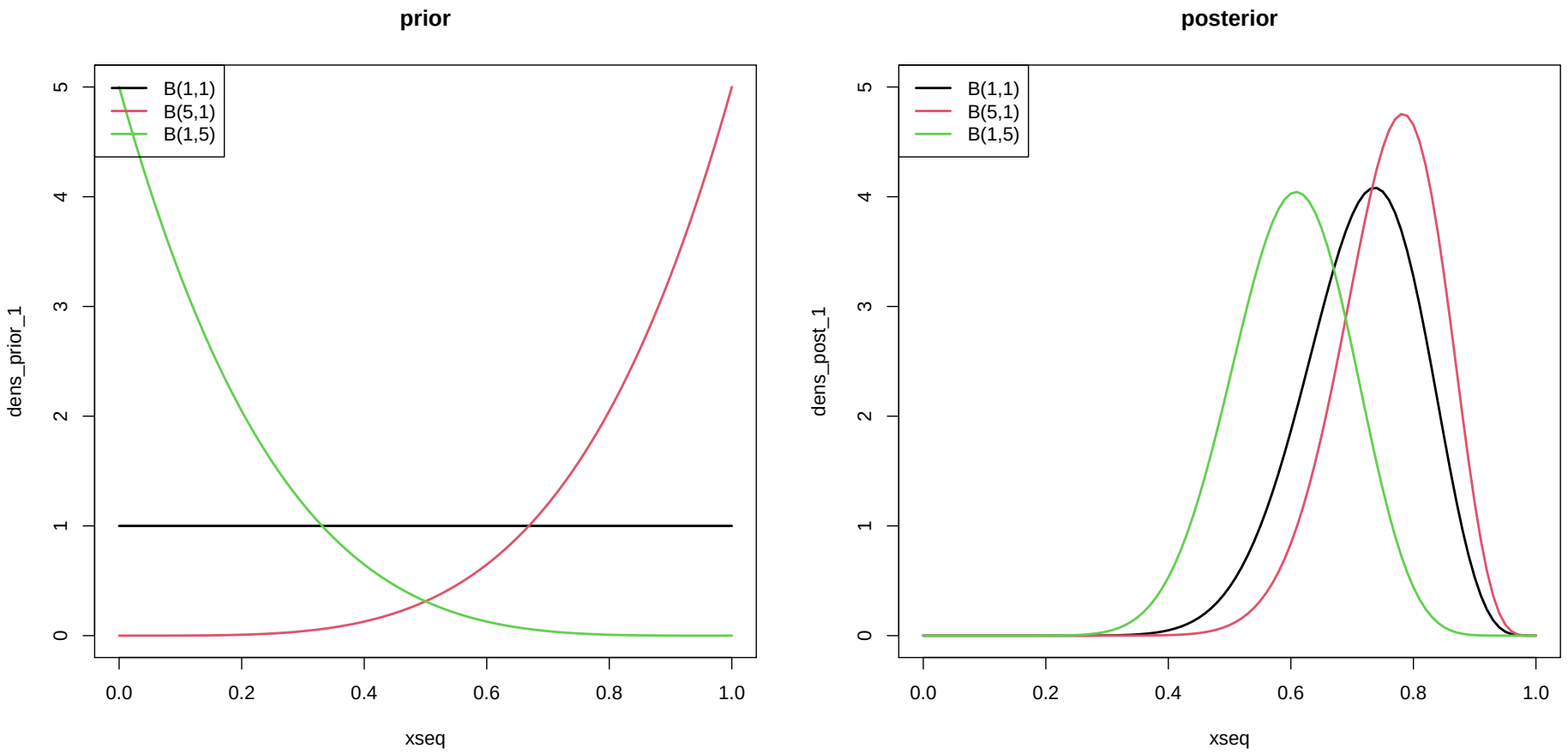


Figure: Posterior dell'esempio con diverse priors, $n = 20$, $y = 15$, e probabilità binomiale vera $= 0.7$.

Code: Codice della figura

```
# settiamo dei parametri e vediamo come cambia
# la posterior rispetto alla prior
# ipotizziamo il valore theta_true come valore vero
set.seed(100)
theta_true = 0.7

# i parametri di 3 prior beta
partheta_a_1 = 1
partheta_b_1 = 1

partheta_a_2 = 5
partheta_b_2 = 1

partheta_a_3 = 1
```

# Il teorema di Bayes

```
partheta_b_3 = 5

# numero di campioni
n = 20

# simuliamo delle osservazioni
y = rbinom(1,n,theta_true)

## plottiamo priori e posteriori
xseq = seq(0,1, by=0.01)

dens_prior_1 = dbeta(xseq,partheta_a_1,partheta_b_1)
dens_prior_2 = dbeta(xseq,partheta_a_2,partheta_b_2)
dens_prior_3 = dbeta(xseq,partheta_a_3,partheta_b_3)
dens_post_1  = dbeta(xseq,y+partheta_a_1-1,n-y+partheta_b_1)
dens_post_2  = dbeta(xseq,y+partheta_a_2-1,n-y+partheta_b_2)
```

# Il teorema di Bayes

```
dens_post_3  = dbeta(xseq,y+partheta_a_3-1,n-y+partheta_b_3)


# priors
#pdf(paste(DIR, "BetaPost.pdf",sep=""), width=7*2)
par(mfrow=c(1,2))
plot(xseq,dens_prior_1, col=1 ,
    ylim=c(0, max(c(dens_prior_1,dens_prior_2,dens_prior_3,
        dens_post_1,dens_post_2,dens_post_3))),
    type="l", lwd=2, main="prior")
lines(xseq,dens_prior_2, col=2, lwd=2)
lines(xseq,dens_prior_3, col=3, lwd=2)
legend("topleft",c("B(1,1)", "B(5,1)","B(1,5)"),
    col=1:3, lty=1, lwd=2)


# posterior
plot(xseq,dens_post_1, col=1 ,
```

# Il teorema di Bayes

```
    ylim=c(0, max(c(dens_prior_1,dens_prior_2,dens_prior_3,
        dens_post_1,dens_post_2,dens_post_3))),
    type="l", lwd=2, main="posterior")
lines(xseq,dens_post_2, col=2, lwd=2)
lines(xseq,dens_post_3, col=3, lwd=2)
legend("topleft",c("B(1,1)", "B(5,1)","B(1,5)"),
col=1:3, lty=1, lwd=2)
```

# I modelli Bayesiani

Per capire la differenza tra statistica Bayesiana e frequentista, prendiamo un semplice esempio: il modello lineare.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

Ricordiamo che stiamo quindi assumendo

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

con $y_i \perp y_j$, e i parametri di interesse sono $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)$, oppure, in forma matriciale

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$$

**APPROCCIO FREQUENTISTA:**

Nell'approccio frequentista, assumiamo i parametri come "fissi" e ignoti.

Calcoliamo la verosimiglianza

$$L(\boldsymbol{\theta}; \mathbf{y}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$$

e il valore di $\boldsymbol{\theta}$ che la massimizza è lo stimatore. La funzione di verosimiglianza ci dice quanto è verosimile che il campione $\mathbf{y}$ sia stato generato da un valore specifico di $\boldsymbol{\theta}$

Essendo lo stimatore funzione delle variabili aleatorie $y_i$, è anch'esso una variabile aleatoria. In altre parole, noi usiamo una variabile aleatoria (la stima) per dire qualcosa su un parametro non aleatorio ($\boldsymbol{\theta}$) ignoto. Nel caso regressivo abbiamo che

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)}{n}$$

In generale, per la varianza si usa

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)}{n-2}$$

dove il $n-2$ deriva dal fatto che abbiamo 2 coefficienti, se fossero $p$, useremmo $n-p$.

**APPROCCIO BAYESIANO:**

Nell'approccio Bayesiano $\boldsymbol{\theta}$ è una variabile aleatoria, la cui distribuzione può rappresenatre

- la reale o presunta legge di probabilità che genera il parametro
- la nostra incertezza circa il valore del parametro stesso (valore deterministico), prima di osservare i dati.

Non ha allora senso massimizzare la verosimiglianza, e quello che possiamo chiederci è, avendo osservato un particolare campione $\mathbf{y}$ come è fatta la distribuzione di $\boldsymbol{\theta}$; vogliamo trovare la distribuzione condizionata

$$f(\boldsymbol{\theta}|\mathbf{y})$$

che secondo il teorema di Bayes è

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{\int f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\mathbf{y})}$$

# I modelli Bayesiani

Lavorare con $f(\boldsymbol{\theta}|\mathbf{y})$ non è facile, dato che generalmente non appartiene a una famiglia di distribuzioni note. Quando $f(\boldsymbol{\theta}|\mathbf{y})$ è una distribuzione nota, si dice che la a priori e la verosimiglianza sono *coniugate* (https://en.wikipedia.org/wiki/Conjugate_prior)

# I modelli Bayesiani

| Likelihood | Model parameters | Conjugate prior distribution | Prior hyperparameters | Posterior hyperparameters[note 1] | Interpretation of hyperparameters | Posterior predictive[note 2] |
|---|---|---|---|---|---|---|
| Bernoulli | $p$ (probability) | Beta | $\alpha,\ \beta \in \mathbb{R}$ | $\alpha + \sum_{i=1}^{n} x_i,\ \beta + n - \sum_{i=1}^{n} x_i$ | $\alpha$ successes, $\beta$ failures[note 3] | $p(\tilde{x}=1) = \frac{\alpha'}{\alpha' + \beta'}$ |
| Binomial with known number of trials, $m$ | $p$ (probability) | Beta | $\alpha,\ \beta \in \mathbb{R}$ | $\alpha + \sum_{i=1}^{n} x_i,\ \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i$ | $\alpha$ successes, $\beta$ failures[note 3] | $\mathrm{BetaBin}(\tilde{x} \mid \alpha', \beta')$ (beta-binomial) |
| Negative binomial with known failure number, $r$ | $p$ (probability) | Beta | $\alpha,\ \beta \in \mathbb{R}$ | $\alpha + rn,\ \beta + \sum_{i=1}^{n} x_i$ | $\alpha$ total successes, $\beta$ failures[note 3] (i.e., $\frac{\beta}{r}$ experiments, assuming $r$ stays fixed) | $\mathrm{BetaNegBin}(\tilde{x} \mid \alpha', \beta')$ (beta-negative binomial) |
| Poisson | $\lambda$ (rate) | Gamma | $k,\ \theta \in \mathbb{R}$ | $k + \sum_{i=1}^{n} x_i,\ \frac{\theta}{n\theta + 1}$ | $k$ total occurrences in $\frac{1}{\theta}$ intervals | $\mathrm{NB}\left(\tilde{x} \mid k', \frac{1}{\theta' + 1}\right)$ (negative binomial) |
| | | | $\alpha,\ \beta$[note 4] | $\alpha + \sum_{i=1}^{n} x_i,\ \beta + n$ | $\alpha$ total occurrences in $\beta$ intervals | $\mathrm{NB}\left(\tilde{x} \mid \alpha', \frac{\beta'}{1 + \beta'}\right)$ (negative binomial) |
| Categorical | $\boldsymbol{p}$ (probability vector), $k$ (number of categories; i.e., size of $\boldsymbol{p}$) | Dirichlet | $\boldsymbol{\alpha} \in \mathbb{R}^k$ | $\boldsymbol{\alpha} + (c_1, \ldots, c_k)$, where $c_i$ is the number of observations in category $i$ | $\alpha_i$ occurrences of category $i$[note 3] | $p(\tilde{x}=i) = \frac{\alpha_i'}{\sum_i \alpha_i'} = \frac{\alpha_i + c_i}{\sum_i \alpha_i + n}$ |
| Multinomial | $\boldsymbol{p}$ (probability vector), $k$ (number of categories; i.e., size of $\boldsymbol{p}$) | Dirichlet | $\boldsymbol{\alpha} \in \mathbb{R}^k$ | $\boldsymbol{\alpha} + \sum_{i=1}^{n} \mathbf{x}_i$ | $\alpha_i$ occurrences of category $i$[note 3] | $\mathrm{DirMult}(\tilde{\mathbf{x}} \mid \boldsymbol{\alpha}')$ (Dirichlet-multinomial) |
| Hypergeometric with known total population size, $N$ | $M$ (number of target members) | Beta-binomial[3] | $n = N, \alpha,\ \beta$ | $\alpha + \sum_{i=1}^{n} x_i,\ \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i$ | $\alpha$ successes, $\beta$ failures[note 3] | |
| Geometric | $p_0$ (probability) | Beta | $\alpha,\ \beta \in \mathbb{R}$ | $\alpha + n,\ \beta + \sum_{i=1}^{n} x_i$ | $\alpha$ experiments, $\beta$ total failures[note 3] | |

Figure: Distribuzioni coniugate

# I modelli Bayesiani

| Likelihood | Model parameters | Conjugate prior distribution | Prior hyperparameters | Posterior hyperparameters[note 1] | Interpretation of hyperparameters | Posterior predictive[note 5] |
|---|---|---|---|---|---|---|
| Normal with known variance $\sigma^2$ | $\mu$ (mean) | Normal | $\mu_0,\ \sigma_0^2$ | $\frac{1}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^n x_i}{\sigma^2}\right),\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}$ | mean was estimated from observations with total precision (sum of all individual precisions) $1/\sigma_0^2$ and with sample mean $\mu_0$ | $\mathcal{N}(\tilde{x}\vert\mu_0', \sigma_0^{2\prime}+\sigma^2)$[4] |
| Normal with known precision $\tau$ | $\mu$ (mean) | Normal | $\mu_0,\ \tau_0^{-1}$ | $\frac{\tau_0\mu_0+\tau\sum_{i=1}^n x_i}{\tau_0+n\tau},\ (\tau_0+n\tau)^{-1}$ | mean was estimated from observations with total precision (sum of all individual precisions)$\tau_0$ and with sample mean $\mu_0$ | $\mathcal{N}\left(\tilde{x}\mid\mu_0', \frac{1}{\tau_0'}+\frac{1}{\tau}\right)$[4] |
| Normal with known mean $\mu$ | $\sigma^2$ (variance) | Inverse gamma | $\alpha,\ \beta$ [note 6] | $\alpha+\frac{n}{2},\ \beta+\frac{\sum_{i=1}^n (x_i-\mu)^2}{2}$ | variance was estimated from $2\alpha$ observations with sample variance $\beta/\alpha$ (i.e. with sum of squared deviations $2\beta$, where deviations are from known mean $\mu$) | $t_{2\alpha'}(\tilde{x}\vert\mu, \sigma^2=\beta'/\alpha')$[4] |
| Normal with known mean $\mu$ | $\sigma^2$ (variance) | Scaled inverse chi-squared | $\nu,\ \sigma_0^2$ | $\nu+n,\ \frac{\nu\sigma_0^2+\sum_{i=1}^n(x_i-\mu)^2}{\nu+n}$ | variance was estimated from $\nu$ observations with sample variance $\sigma_0^2$ | $t_{\nu'}(\tilde{x}\vert\mu, \sigma_0^{2\prime})$[4] |
| Normal with known mean $\mu$ | $\tau$ (precision) | Gamma | $\alpha,\ \beta$[note 4] | $\alpha+\frac{n}{2},\ \beta+\frac{\sum_{i=1}^n(x_i-\mu)^2}{2}$ | precision was estimated from $2\alpha$ observations with sample variance $\beta/\alpha$ (i.e. with sum of squared deviations $2\beta$, where deviations are from known mean $\mu$) | $t_{2\alpha'}(\tilde{x}\mid\mu, \sigma^2=\beta'/\alpha')$[4] |
| Normal[note 7] | $\mu$ and $\sigma^2$ Assuming exchangeability | Normal-inverse gamma | $\mu_0,\ \nu,\ \alpha,\ \beta$ | $\frac{\nu\mu_0+n\bar{x}}{\nu+n},\ \nu+n,\ \alpha+\frac{n}{2},$ $\beta+\frac{1}{2}\sum_{i=1}^n(x_i-\bar{x})^2+\frac{n\nu}{\nu+n}\frac{(\bar{x}-\mu_0)^2}{2}$ • $\bar{x}$ is the sample mean | mean was estimated from $\nu$ observations with sample mean $\mu_0$; variance was estimated from $2\alpha$ observations with sample mean $\mu_0$ and sum of squared deviations $2\beta$ | $t_{2\alpha'}\left(\tilde{x}\mid\mu', \frac{\beta'(\nu'+1)}{\nu'\alpha'}\right)$[4] |
| Normal | $\mu$ and $\tau$ Assuming exchangeability | Normal-gamma | $\mu_0,\ \nu,\ \alpha,\ \beta$ | $\frac{\nu\mu_0+n\bar{x}}{\nu+n},\ \nu+n,\ \alpha+\frac{n}{2},$ $\beta+\frac{1}{2}\sum_{i=1}^n(x_i-\bar{x})^2+\frac{n\nu}{\nu+n}\frac{(\bar{x}-\mu_0)^2}{2}$ • $\bar{x}$ is the sample mean | mean was estimated from $\nu$ observations with sample mean $\mu_0$, and precision was estimated from $2\alpha$ observations with sample mean $\mu_0$ and sum of squared deviations $2\beta$ | $t_{2\alpha'}\left(\tilde{x}\mid\mu', \frac{\beta'(\nu'+1)}{\alpha'\nu'}\right)$[4] |

Figure: Distribuzioni coniugate

# I modelli Bayesiani

| | | | | | | |
|---|---|---|---|---|---|---|
| Multivariate normal with known covariance matrix $\boldsymbol{\Sigma}$ | $\boldsymbol{\mu}$ (mean vector) | Multivariate normal | $\boldsymbol{\mu}_0,\ \boldsymbol{\Sigma}_0$ | $\left(\boldsymbol{\Sigma}_0^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + n\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}\right),$ $\left(\boldsymbol{\Sigma}_0^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1}$ • $\bar{\mathbf{x}}$ is the sample mean | mean was estimated from observations with total precision (sum of all individual precisions) $\boldsymbol{\Sigma}_0^{-1}$ and with sample mean $\boldsymbol{\mu}_0$ | $\mathcal{N}(\tilde{\mathbf{x}} \mid \boldsymbol{\mu}_0{}', \boldsymbol{\Sigma}_0{}' + \boldsymbol{\Sigma})$[4] |
| Multivariate normal with known precision matrix $\boldsymbol{\Lambda}$ | $\boldsymbol{\mu}$ (mean vector) | Multivariate normal | $\boldsymbol{\mu}_0,\ \boldsymbol{\Lambda}_0$ | $\left(\boldsymbol{\Lambda}_0 + n\boldsymbol{\Lambda}\right)^{-1}\left(\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 + n\boldsymbol{\Lambda}\bar{\mathbf{x}}\right),\ \left(\boldsymbol{\Lambda}_0 + n\boldsymbol{\Lambda}\right)$ • $\bar{\mathbf{x}}$ is the sample mean | mean was estimated from observations with total precision (sum of all individual precisions) $\boldsymbol{\Lambda}_0$ and with sample mean $\boldsymbol{\mu}_0$ | $\mathcal{N}\left(\tilde{\mathbf{x}} \mid \boldsymbol{\mu}_0{}', \boldsymbol{\Lambda}_0{}'^{-1} + \boldsymbol{\Lambda}^{-1}\right)$[4] |
| Multivariate normal with known mean $\boldsymbol{\mu}$ | $\boldsymbol{\Sigma}$ (covariance matrix) | Inverse-Wishart | $\nu,\ \boldsymbol{\Psi}$ | $n+\nu,\ \boldsymbol{\Psi} + \sum_{i=1}^{n}(\mathbf{x_i} - \boldsymbol{\mu})(\mathbf{x_i} - \boldsymbol{\mu})^T$ | covariance matrix was estimated from $\nu$ observations with sum of pairwise deviation products $\boldsymbol{\Psi}$ | $t_{\nu'-p+1}\left(\tilde{\mathbf{x}}\middle\vert\boldsymbol{\mu}, \frac{1}{\nu'-p+1}\boldsymbol{\Psi}'\right)$[4] |
| Multivariate normal with known mean $\boldsymbol{\mu}$ | $\boldsymbol{\Lambda}$ (precision matrix) | Wishart | $\nu,\ \mathbf{V}$ | $n+\nu,\ \left(\mathbf{V}^{-1} + \sum_{i=1}^{n}(\mathbf{x_i} - \boldsymbol{\mu})(\mathbf{x_i} - \boldsymbol{\mu})^T\right)^{-1}$ | covariance matrix was estimated from $\nu$ observations with sum of pairwise deviation products $\mathbf{V}^{-1}$ | $t_{\nu'-p+1}\left(\tilde{\mathbf{x}} \mid \boldsymbol{\mu}, \frac{1}{\nu'-p+1}\mathbf{V}'^{-1}\right)$[4] |
| Multivariate normal | $\boldsymbol{\mu}$ (mean vector) and $\boldsymbol{\Sigma}$ (covariance matrix) | normal-inverse-Wishart | $\boldsymbol{\mu}_0,\ \kappa_0,\ \nu_0,\ \boldsymbol{\Psi}$ | $\frac{\kappa_0\boldsymbol{\mu}_0 + n\bar{\mathbf{x}}}{\kappa_0 + n},\ \kappa_0 + n,\ \nu_0 + n,$ $\boldsymbol{\Psi} + \mathbf{C} + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^T$ • $\bar{\mathbf{x}}$ is the sample mean • $\mathbf{C} = \sum_{i=1}^{n}(\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})^T$ | mean was estimated from $\kappa_0$ observations with sample mean $\boldsymbol{\mu}_0$; covariance matrix was estimated from $\nu_0$ observations with sample mean $\boldsymbol{\mu}_0$ and with sum of pairwise deviation products $\boldsymbol{\Psi} = \nu_0\boldsymbol{\Sigma}_0$ | $t_{\nu_0'-p+1}\left(\tilde{\mathbf{x}}\middle\vert\boldsymbol{\mu}_0{}', \frac{\kappa_0{}'+1}{\kappa_0{}'(\nu_0{}'-p+1)}\boldsymbol{\Psi}'\right)$[4] |
| Multivariate normal | $\boldsymbol{\mu}$ (mean vector) and $\boldsymbol{\Lambda}$ (precision matrix) | normal-Wishart | $\boldsymbol{\mu}_0,\ \kappa_0,\ \nu_0,\ \mathbf{V}$ | $\frac{\kappa_0\boldsymbol{\mu}_0 + n\bar{\mathbf{x}}}{\kappa_0 + n},\ \kappa_0 + n,\ \nu_0 + n,$ $\left(\mathbf{V}^{-1} + \mathbf{C} + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^T\right)^{-1}$ • $\bar{\mathbf{x}}$ is the sample mean • $\mathbf{C} = \sum_{i=1}^{n}(\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})^T$ | mean was estimated from $\kappa_0$ observations with sample mean $\boldsymbol{\mu}_0$; covariance matrix was estimated from $\nu_0$ observations with sample mean $\boldsymbol{\mu}_0$ and with sum of pairwise deviation products $\mathbf{V}^{-1}$ | $t_{\nu_0'-p+1}\left(\tilde{\mathbf{x}} \mid \boldsymbol{\mu}_0{}', \frac{\kappa_0{}'+1}{\kappa_0{}'(\nu_0{}'-p+1)}\mathbf{V}'^{-1}\right)$ [4] |
| Uniform | $U(0,\theta)$ | Pareto | $x_m,\ k$ | $\max\{x_1,\ldots,x_n,x_m\},\ k+n$ | $k$ observations with maximum value $x_m$ | |

Figure: Distribuzioni coniugate

# I modelli Bayesiani

| | | | | | | |
|---|---|---|---|---|---|---|
| Pareto with known minimum $x_m$ | $k$ (shape) | Gamma | $\alpha,\ \beta$ | $\alpha + n,\ \beta + \sum_{i=1}^{n} \ln \frac{x_i}{x_\mathrm{m}}$ | $\alpha$ observations with sum $\beta$ of the order of magnitude of each observation (i.e. the logarithm of the ratio of each observation to the minimum $x_m$) | |
| Weibull with known shape $\beta$ | $\theta$ (scale) | Inverse gamma[3] | $a, b$ | $a + n,\ b + \sum_{i=1}^{n} x_i^{\beta}$ | $a$ observations with sum $b$ of the $\beta$th power of each observation | |
| Log-normal | Same as for the normal distribution after applying the natural logarithm to the data for the posterior hyperparameters. Please refer to Fink (1997, pp. 21–22) to see the details. | | | | | |
| Exponential | $\lambda$ (rate) | Gamma | $\alpha,\ \beta$[note 4] | $\alpha + n,\ \beta + \sum_{i=1}^{n} x_i$ | $\alpha - 1$ observations that sum to $\beta$ [5] | $\mathrm{Lomax}(\tilde{x} \mid \beta', \alpha')$ (Lomax distribution) |
| Gamma with known shape $\alpha$ | $\beta$ (rate) | Gamma | $\alpha_0,\ \beta_0$ | $\alpha_0 + n\alpha,\ \beta_0 + \sum_{i=1}^{n} x_i$ | $\alpha_0/\alpha$ observations with sum $\beta_0$ | $\mathrm{CG}(\tilde{\mathbf{x}} \mid \alpha, \alpha_0', \beta_0') = \beta'(\tilde{\mathbf{x}} \mid \alpha, \alpha_0', 1, \beta_0')$ [note 8] |
| Inverse Gamma with known shape $\alpha$ | $\beta$ (inverse scale) | Gamma | $\alpha_0,\ \beta_0$ | $\alpha_0 + n\alpha,\ \beta_0 + \sum_{i=1}^{n} \frac{1}{x_i}$ | $\alpha_0/\alpha$ observations with sum $\beta_0$ | |
| Gamma with known rate $\beta$ | $\alpha$ (shape) | $\propto \frac{a^{\alpha-1}\beta^{\alpha c}}{\Gamma(\alpha)^b}$ | $a,\ b,\ c$ | $a \prod_{i=1}^{n} x_i,\ b + n,\ c + n$ | $b$ or $c$ observations ($b$ for estimating $\alpha$, $c$ for estimating $\beta$) with product $a$ | |
| Gamma[3] | $\alpha$ (shape), $\beta$ (inverse scale) | $\propto \frac{p^{\alpha-1}e^{-\beta q}}{\Gamma(\alpha)^r \beta^{-\alpha s}}$ | $p,\ q,\ r,\ s$ | $p \prod_{i=1}^{n} x_i,\ q + \sum_{i=1}^{n} x_i,\ r + n,\ s + n$ | $\alpha$ was estimated from $r$ observations with product $p$; $\beta$ was estimated from $s$ observations with sum $q$ | |
| Beta | $\alpha, \beta$ | $\propto \frac{\Gamma(\alpha+\beta)^k\, p^{\alpha}\, q^{\beta}}{\Gamma(\alpha)^k\, \Gamma(\beta)^k}$ | $p,\ q,\ k$ | $p \prod_{i=1}^{n} x_i,\ q \prod_{i=1}^{n} (1 - x_i),\ k + n$ | $\alpha$ and $\beta$ were estimated from $k$ observations with product $p$ and product of the complements $q$ | |

Figure: Distribuzioni coniugate

## Kernel e costante di normalizzazione

Per capire e implementare i modelli Bayesiani, è fondamentale capire il concetto di kernel e costante di normalizzazione. Ipotizziamo di star lavorando con una densità $f(\mathbf{x}|\mathbf{y})$ generale, dove $\mathbf{y}$ potrebbero essere parametri o no.

Attenzione: ricordate che nel Bayesiano non c'è differenza tra parametri e osservazioni, nel senso che sono entrambe variabili aleatorie

La densità $f(\mathbf{x}|\mathbf{y})$ si può dividere in due parti

$$f(\mathbf{x}|\mathbf{y}) = \frac{k(\mathbf{x}|\mathbf{y})}{C(\mathbf{y})}$$

dove $k(\mathbf{x}|\mathbf{y})$, chiamato kernel, che dipende dalla variabile aleatoria alla sinistra (quella di cui stiamo calcolando la densità), e una costante di normalizzazione $C(\mathbf{y})$ che non

dipende dalla variabile aleatoria $(\mathbf{x})$. Una distribuzione è totalmente descritta dal solo kernel visto che

$$1 = \int_{\mathcal{X}} f(\mathbf{x}|\mathbf{y})d\mathbf{x} = \int_{\mathcal{X}} \frac{k(\mathbf{x}|\mathbf{y})}{C(\mathbf{y})}d\mathbf{x} = \frac{1}{C(\mathbf{y})}\int_{\mathcal{X}} k(\mathbf{x}|\mathbf{y})d\mathbf{x}$$

e quindi deve essere

$$\int_{\mathcal{X}} k(\mathbf{x}|\mathbf{y})d\mathbf{x} = C(\mathbf{y})$$

e

$$f(\mathbf{x}|\mathbf{y}) \propto k(\mathbf{x}|\mathbf{y})$$

vediamo alcuni kernel:

- $X \sim G(a, b)$, $f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$,

$$k(x|a, b) = x^{a-1} \exp(-bx)$$

# I modelli Bayesiani

- $X \sim IG(a,b)$, $f(x) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp\left(-\frac{b}{x}\right)$,

$$k(x|a,b) = x^{-a-1} \exp\left(-\frac{b}{x}\right)$$

- $X \sim Beta(a,b)$, $f(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$,

$$k(x|a,b) = x^{a-1}(1-x)^{b-1}$$

- $X \sim Bin(n,p)$, $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$,

$$k(x|n,p) = \frac{1}{x!(n-x)!} p^x (1-p)^{n-x}$$

- $X \sim N(\mu,\sigma^2)$, $f(x) = (2\pi\sigma^2)^{-0.5} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$,

$$k(x|\mu,\sigma^2) = \exp\left(-\frac{x^2 - 2x\mu}{2\sigma^2}\right)$$

- $\mathbf{X} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $f(x) = (2\pi)^{-\frac{m}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}{2}\right)$,

$$k(x|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \exp\left(-\frac{\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} - 2\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}}{2}\right)$$

L'importanza del kernel si può capire riprendendo l'esempio binomiale. Ricordiamo che la a posteriori è

$$f(\theta|y) = \frac{f(y|\theta) f(\theta)}{f(y)}$$

ma in questo caso $f(y)$ è la costante di normalizzazione di $f(\theta|y)$, e $f(y|\theta) f(\theta)$ il kernel. Dobbiamo quindi solo calcolare il kernel

$$f(\theta|y) \propto f(y|\theta) f(\theta)$$

e vedere se riusciamo a ricondurlo a qualche distribuzioni conosciuto. Nel caso di prima abbiamo che

$$f(y|\theta)f(\theta) = \left( \begin{array}{c} n \\ y \end{array} \right) \frac{\theta^{y+a-1}(1-\theta)^{n+b-y-1}}{B(a,b)} \propto \theta^{y+a-1}(1-\theta)^{n+b-y-1}$$

che possiamo immediatamente riconoscere come il kernel di una beta di parametri $y+a$ e $n+b-y$, senza dover calcolare il denominatore.

Il kernel ci sarà utile quando introdurremo gli algoritmo **MCMC** (Markov Chain Monte Carlo)

Prendiamo un'altro esempio di distribuzioni coniugate:

$$\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$
$$\boldsymbol{\beta} \sim N_p(\mathbf{M}, \mathbf{V}).$$

assumendo $\sigma^2$ noto. Noi siamo interessati alla posteriori $f(\boldsymbol{\beta}|\mathbf{y})$
Abbiamo che

$$f(\mathbf{y}|\boldsymbol{\beta}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right)$$
$$f(\boldsymbol{\beta}) = (2\pi)^{-\frac{p}{2}}|\mathbf{V}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{M})^T\mathbf{V}^{-1}(\boldsymbol{\beta}-\mathbf{M})\right)$$

I calcoli si semplificano ricordando che per trovare la distribuzione possiamo guardare solo al kernel (tutto ciò che dipende dalla variabile aleatorie). In questo caso la variabile aleatoria è $\boldsymbol{\beta}$, dato che stiamo cercando la sua a posteriori. Abbiamo quindi che

$$f(\boldsymbol{\beta}|\mathbf{y}) \propto \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}\right)\right) \times$$
$$\exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}^T\mathbf{V}^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbf{V}^{-1}\mathbf{M}\right)\right) =$$
$$\exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}^T(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \mathbf{V}^{-1})\boldsymbol{\beta} - 2\boldsymbol{\beta}^T(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{y} + \mathbf{V}^{-1}\mathbf{M})\right)\right)$$

Se definiamo

$$\mathbf{V}_p = \left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \mathbf{V}^{-1}\right)^{-1}$$
$$\mathbf{M}_p = \mathbf{V}_p\left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{y} + \mathbf{V}^{-1}\mathbf{M}\right)$$

abbiamo che

$$f(\boldsymbol{\beta}|\mathbf{y}) \propto \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}^T\mathbf{V}_p^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbf{V}_p^{-1}\mathbf{M}_p\right)\right)$$

che è il kernel di una normale multivariata, quindi

$$\boldsymbol{\beta}|\mathbf{y} \sim N_p(\mathbf{M}_p, \mathbf{V}_p)$$

Attenzione anche in questo semplice esempio, siamo in grado di scrivere la a posteriori in forma chiusa, solo perchè abbiamo assunto $\sigma^2$ nota, altrimenti, non sarebbe stato possibile.

Anche se il modello Bayesiano e Frequentista sono diversi, non ci aspettiamo di trovare grandi differenze, soprattutto se i dati sono molti e le prior sono "poco informative".
Per esempio. Sappiamo che lo stimatore di massima verosimiglianza di $\boldsymbol{\beta}$ è

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

che è distribuito normalmente Adesso, supponiamo che $\mathbf{V}$ sia diagonale e con varianze molto elevate. Allora $\mathbf{V}^{-1} \approx 0\mathbf{I}$ e quindi

$$\mathbf{V}_p \approx (\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X})^{-1}$$

$$\mathbf{M}_p \approx \mathbf{V}_p(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{y}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

quindi la media della a posteriori è uguale allo stimatore di massima verosimiglianza.

Il risultato non è certo una coincidenza, e per capirne il motivo, riprendiamo la a posteriori

$$f(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)f(\theta)}{f(\mathbf{y})}$$

se prendiamo una distribuzione a priori $f(\theta)$ che è approssimativamente costante

$$f(\theta) \approx c$$

allora

$$f(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)c}{f(\mathbf{y})} \propto f(\mathbf{y}|\theta)$$

Notate che se $\theta$ è continua e il suo dominio non limitato, $f(\theta)$ non può essere costante.
Sappiamo anche che

$$f(\mathbf{y}|\theta) = L(\theta|\mathbf{y})$$

e quindi la a-posteriori è proporzionale alla verosimiglianza e i massimi delle due corrispondono. Ricordate che $L(\theta|\mathbf{y})$ non è una densità, e questo lo si vede dalla formula precedente, visto che per diventare una densità deve essere moltiplicata per $c/f(\mathbf{y})$.
In altre parole, la verosimiglianza può essere vista come il kernel della a posteriori quando le a priori sono non informative.

Prima di passare ai casi più generali, concludiamo con un semplice esempio che mostra come il modello Bayesiano è un modo generale per fare update dell'informazione ogni volta che si fa un esperimento.

# I modelli Bayesiani

Ipotizziamo che il parametro di interesse sia $\theta$ e che abbiamo osservato un campione $\mathbf{x}_1$ da $f(\mathbf{x}_1|\theta)$. Siamo allora interessati alla a-posteriori

$$f(\theta|\mathbf{x}_1) = \frac{f(\mathbf{x}_1|\theta)f(\theta)}{f(\mathbf{x}_1)}$$

Se invece di $\mathbf{x}_1$ avessimo osservato $\mathbf{x}_1$ e $\mathbf{x}_2$, allora la a posteriori di riferimento sarebbe

$$f(\theta|\mathbf{x}_1, \mathbf{x}_2) = \frac{f(\mathbf{x}_1, \mathbf{x}_2|\theta)f(\theta)}{f(\mathbf{x}_1, \mathbf{x}_2)} = \frac{f(\mathbf{x}_2|\mathbf{x}_1\theta)f(\mathbf{x}_1|\theta)f(\theta)}{f(\mathbf{x}_2|\mathbf{x}_1)f(\mathbf{x}_1)} = \frac{f(\mathbf{x}_2|\mathbf{x}_1\theta)f(\theta|\mathbf{x}_1)}{f(\mathbf{x}_2|\mathbf{x}_1)}$$

quindi la a-posteriori $f(\theta|\mathbf{x}_1)$ può essere vista come la nuova a-priori se osserviamo un nuovo dato $\mathbf{x}_2$.

Più schematicamente:

- osservo $\mathbf{x}_1$
- la mia informazione a priori sul parametro è $f(\theta)$
- tramite la verosimiglianza $f(\mathbf{x}_1|\theta)$ passo dalla a-priori $f(\theta)$ alla a-posteriori $f(\theta|\mathbf{x}_1)$

# I modelli Bayesiani

- osservo $\mathbf{x}_2$
- la mia nuova informazione a priori sul parametro è data da $f(\theta|\mathbf{x}_1)$
- utilizzando la verosimiglianza $f(\mathbf{x}_2|\mathbf{x}_1\theta)$ passo dalla a priori $f(\theta|\mathbf{x}_1)$ alla a-posteriori $f(\theta|\mathbf{x}_1, \mathbf{x}_2)$
- $f(\theta|\mathbf{x}_1, \mathbf{x}_2)$ può essere vista come la a priori su $\theta$ per futuri esperimenti