Modelli a parametro singolo

Modello binomiale

Docente: Matilde Trevisani

DEAMS

A.A. 2025/2026
(aggiornato: 2025-10-10)

Agenda

Modelli a parametro singolo

- Modello Binomiale
- Nota sull'effetto di più evidenza
- Riassumere l'inferenza a posteriori
- Coniugatezza
- *Interplay* tra a priori e dati

Modelli a parametro singolo

Modello binomiale

Binomial model: a one-parameter model

We start to illustrate Bayesian inference in the context of statistical models where only a single scalar parameter is to be estimated; that is, the estimand θ is **onedimensional**.

We start with the Binomial model where the aim is estimating a probability from binomial data, i.e., the results of a sequence of 'Bernoulli trials'.

Although a very simple model, it has relevant applications.

It is also used as a building block in more complex models.

Also, it was dealt with by many of the first scholars working in probability.

In fact, it was the motivating example to develop Bayesian statistics both for T. Bayes and for Laplace. The former considered it in an abstract context, the latter had the aim of estimating the probability of a female birth.

Binomial data

We observe the results of a sequence of 'Bernoulli trials' (trials or draws from a large population), i.e., data $y_1, \ldots, y_n$ each of which is either 0 or 1 (coding 'failure' and 'success' labels, respectively).

Let θ be the probability of success in each trial. If we consider the trials *exchangeable*, that is equivalent to say that, conditional on θ, the $y_1, \ldots, y_n$ are iid, i.e.,

-- independent: if $i \neq j$, $P(y_i = 1|y_j = 1, \theta) = P(y_i = 1|\theta)$
-- identically distributed: $P(y_i = 1|\theta) = \theta \ \ \forall i$.
so that, e.g., a sequence $1, 1, 0, 1, 0, 0, \ldots$ has probability $\theta\theta(1-\theta)\theta(1-\theta)(1-\theta)\ldots$

we are saying that **we disregard the order** and the data can be summarized by the total number of 1 (successes), which we denote by y.

If n is the # of trials, the sampling model for $y|\theta$ is then a binomial model

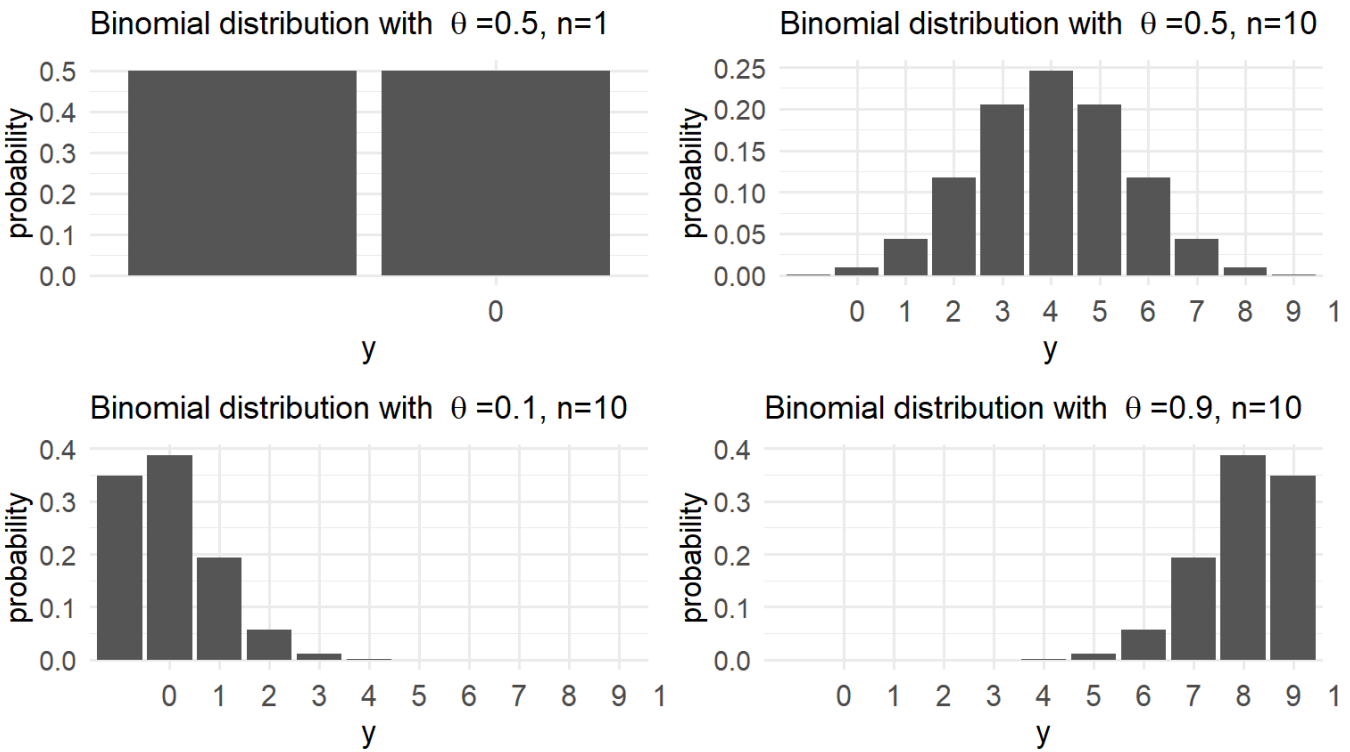
$$p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

where on the left side we suppress the dependence on n because it is regarded as part of the experimental design that is considered fixed(; all the probabilities discussed for this problem are assumed to be conditional on n).

Binomial model $(y|\theta)$: θ known

- **Observational model**(/sampling distribution/statistical model) (discrete function of y)

$$p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

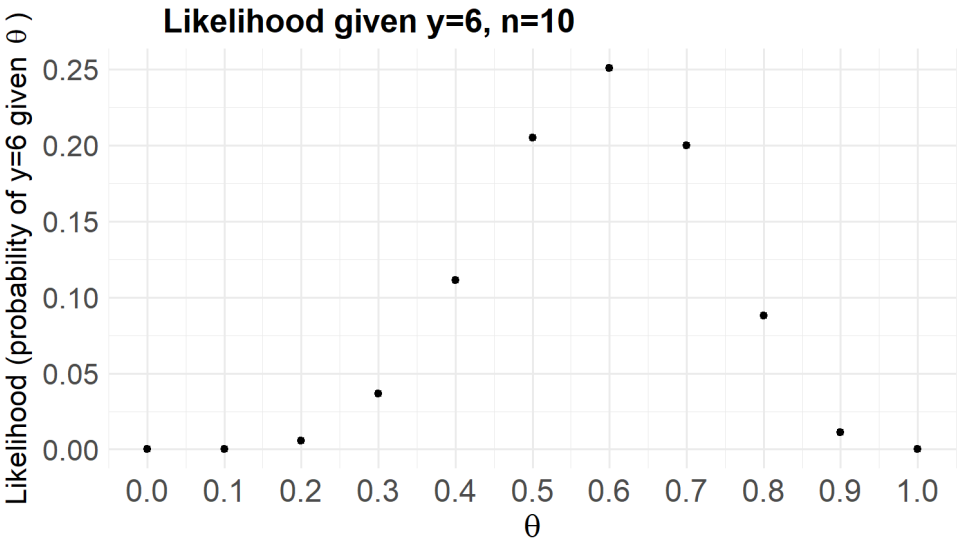


Binomial model as likelihood: θ unknown

Likelihood (continuous function of θ)

$$p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

E.g., consider $y = 6$ and $n = 10$



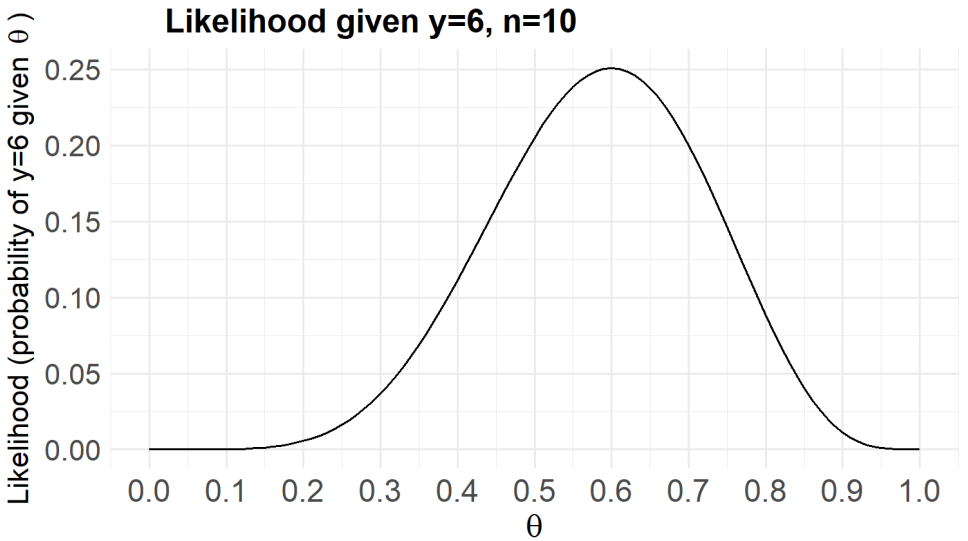
$p(y = 6|n = 10, \theta)$: 0.00 0.00 0.01 0.04 0.11 0.21 0.25 0.20 0.09 0.01 0.00

Binomial model as likelihood: θ unknown

Likelihood (continuous function of θ)

$$p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

E.g., consider $y = 6$ and $n = 10$



`integrate(function(`θ`) dbinom(6, 10,`θ`), 0, 1)` $\approx 0.09 \neq 1$

Binomial model with uniform prior

The **posterior** (continous function of θ) by the *Bayes rule*

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

where $p(y) = \int p(y|\theta)p(\theta)d\theta$

Let's start with a *uniform* prior

$$p(\theta) = 1, \text{ with } 0 \leq \theta \leq 1$$

Hence

$$\begin{aligned} p(\theta|y) = \frac{p(y|\theta)}{p(y)} &= \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}}{\int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}d\theta} \\ &= \frac{1}{Z}\theta^y(1-\theta)^{n-y} \end{aligned}$$

Binomial model with uniform prior

- Z (costante dato y) è il termine di normalizzazione

$$Z = \int_0^1 \theta^y (1-\theta)^{n-y} d\theta = \frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)}$$

- Il termine di normalizzazione ha la forma della funzione Beta
  - quando integrato su $(0, 1)$ il risultato può essere presentato con Funzioni gamma
  - con numeri interi $\Gamma(n) = (n-1)!$
  - se i numeri interi sono grandi anche questo calcolo è impegnativo e di solito $\log \Gamma(\cdot)$ viene calcolato invece di $\Gamma(\cdot)$

- Valutiamolo con $y = 6, n = 10$

`y<-6;n<-10; integrate(function(theta) theta^y*(1-theta)^(n-y), 0, 1)` ≈ 0.0004329

`gamma(y+1)*gamma(n-y+1)/gamma(n+2)` ≈ 0.0004329

Binomial model with uniform prior

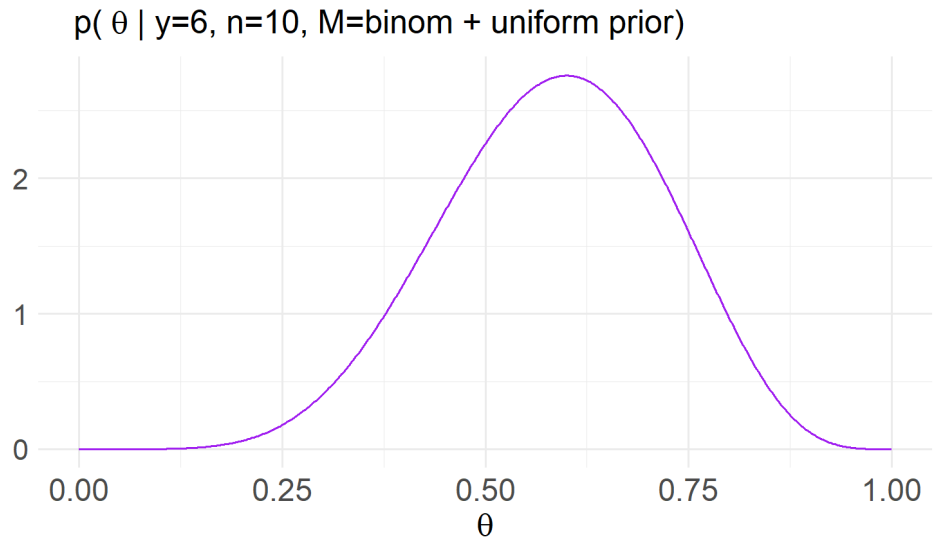
The **posterior**

$$p(\theta|y) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^y(1-\theta)^{n-y},$$

is the *Beta* distribution with parameters $y+1$ and $n-y+1$, and we can also write

$$\theta|y \sim \text{Beta}(y+1, n-y+1)$$

E.g., consider $y = 6$ and $n = 10$



Conditioning to the model

A volte il condizionamento al modello M viene mostrato esplicitamente

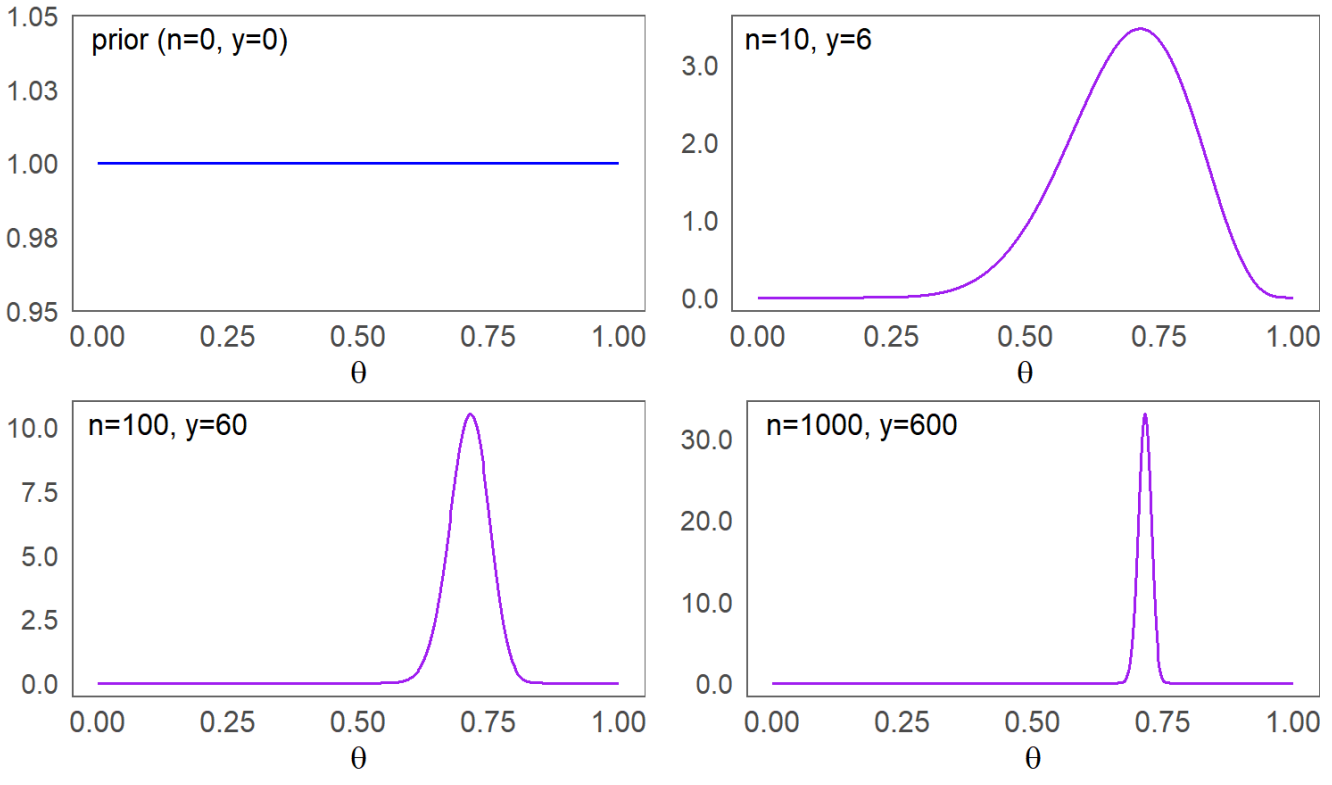
La a posteriori con la regola di Bayes (funzione di θ, continua)

$$p(\theta|y, M) = \frac{p(y|\theta, M)p(\theta|M)}{p(y|M)}$$

dove $p(y|M) = \int p(y|\theta, M)p(\theta|M)d\theta$

- rende più chiaro che la verosimiglianza e la a priori fanno entrambe parte del modello
- rende più chiaro che non esiste una probabilità assoluta per $p(y)$, ma essa dipende dal modello M
- in caso di due modelli, possiamo valutare le probabilità marginali $p(y|M_1)$ e $p(y|M_2)$
- di solito è sottintesa per rendere la notazione più concisa

Posterior densities for binomial parameter θ



Posterior density for binomial parameter θ, based on uniform prior distribution and y successes out of n trials. Curves displayed for several values of n and y.

Still on Bayes' Thorem: Impact of More Evidence

Incorporating Evidence

- Reasoning under uncertainty requires a mechanism for incorporating evidence
- Bayes' theorem as an updating mechanism
  - From prior to posterior
- Properly synthesizes information in the data to revise the probability distribution for the unknown parameter

1. **Impact of More Evidence**
   The more data we have, the more the posterior reflects that
   - As sample size increases, the posterior becomes increasing similar to the likelihood (usually)
2. **Accumulation of Evidence**
   As new data arrives, proper synthesis, updating of the distribution
   - Today's posterior is tomorrow's prior

Laplace example, revisited

Laplace observed $241\,945$ females and $251\,527$ males, that is if

$$\theta = \text{probability of a female birth}$$

he had

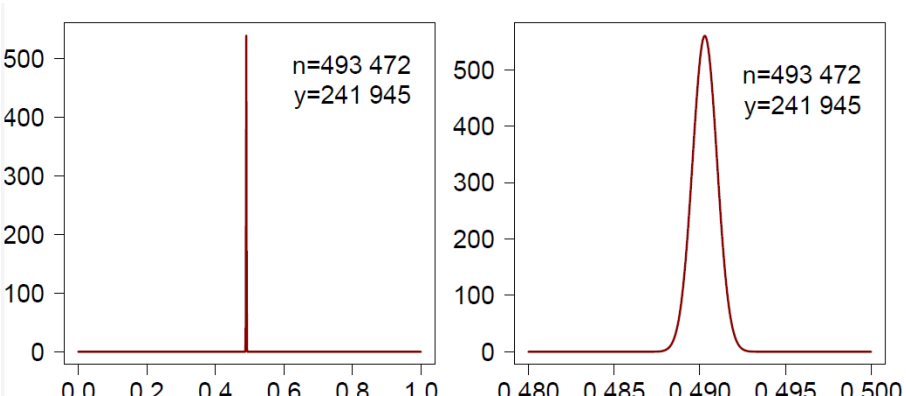
$$n = 241\,945 + 251\,527 = 493\,472; \quad y = 241\,945$$

hence the posterior distribution for θ is a $\text{Beta}(241\,946, 251\,528)$ and

$$P(\theta \geq 0.5|y) \approx 1.15 \times 10^{-42}$$

We ought to appreciate the fact that to get to this number Laplace had to develop appropriate approximations, it is not immediate even today (R may give 0 depending on how the problem is formulated due to machine precision).

Posterior distribution for Laplace

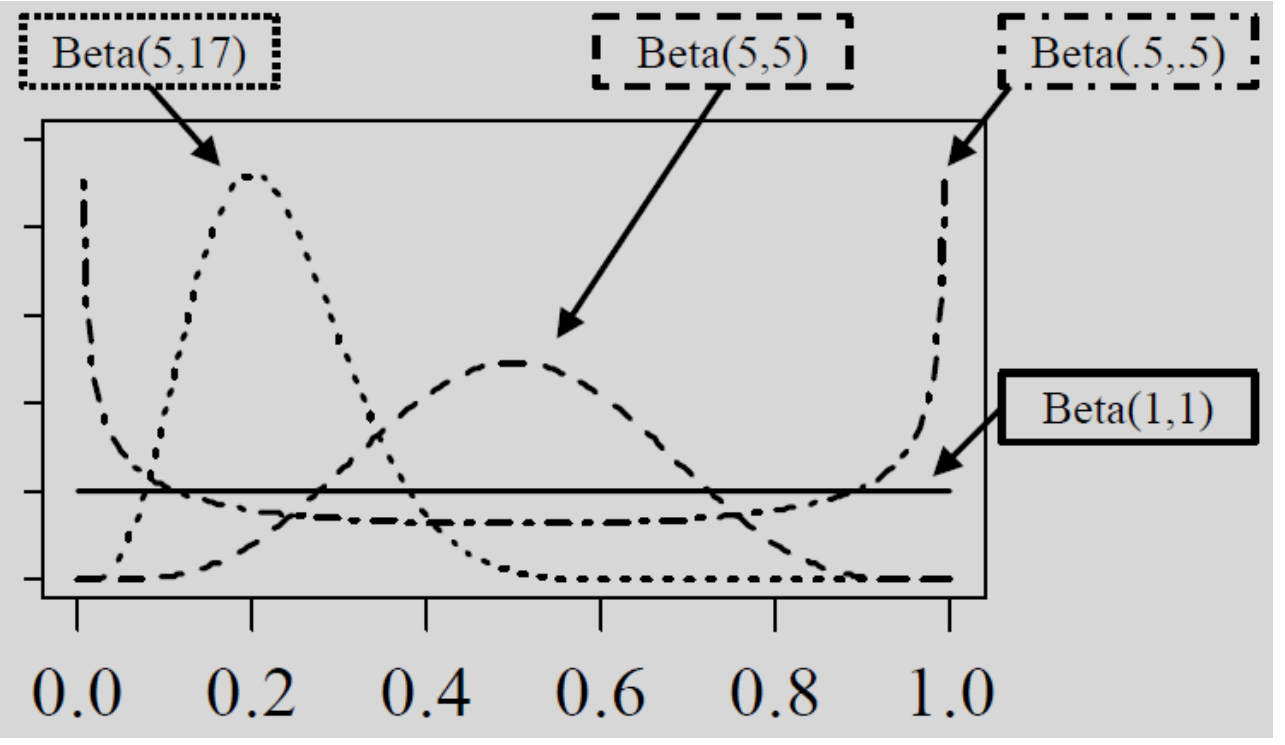


Modello binomiale: calcolo

- R
  - density `dbeta`
  - CDF `pbeta`
  - quantile `qbeta`
  - random number `rbeta`

- Beta CDF non banale calcolarla
- E.g., `pbeta` in `R` usa una frazione continua con fattori di ponderazione e espansione asintotica
- Bayes was able to solve integral given small n and y. In case of large n and y, Laplace developed a Gaussian approximation (*Laplace approximation*) of the posterior. In this specific case, R `pbeta` gives the same results as Laplace's result with at least 3 digit accuracy.

Distribuzioni Beta



Distribuzioni Beta

- Distribuzione su $[0, 1]$
- $\theta|\alpha, \beta \sim Beta(\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$
- La costante di normalizzazione è il reciproco di

$$Z = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

- La densità è finita se $\alpha, \beta \geq 1$, e l'integrale è finito se se $\alpha, \beta > 0$.
  - La densità ha un asintoto in 0 se $\alpha < 1$, in 1 se $\beta < 1$.
- $E(Beta(\alpha, \beta)) = \frac{\alpha}{\alpha+\beta}$
- $\text{Var}(Beta(\alpha, \beta)) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
- $\text{Moda}(Beta(\alpha, \beta)) = \frac{\alpha-1}{\alpha+\beta-2}$ se $\alpha, \beta > 1$, $= 1$ se $\alpha \geq 1, \beta < 1$, $= 0$ se $\alpha < 1, \beta \geq 1$; ha due mode se $\alpha, \beta < 1$
- $\alpha = E(\theta)(\alpha+\beta)$, $\beta = (1-E(\theta))(\alpha+\beta)$

Inferenza con dati binomiali

Approccio classico/frequentista

Dato θ, quali sono le probabilità dei vari possibili risultati per la v.a. y?

Legge (debole) dei grandi numeri (*Weak Law of Large Numbers*) (teorema di Bernoulli)

$$y \sim \text{Bin}(n, \theta)$$

$$lim_{n \to \infty} P\left(\frac{y}{n} - \theta > \epsilon \,|\, \theta\right) = 0$$

- MLE: $\hat{\theta} = \frac{y}{n}$

Approccio Bayesiano

Dato y, quali sono le probabilità dei vari possibili risultati per la v.a. θ?

- $[\theta|y]$
- as well as summaries of the posterior distribution

Summarizing posterior distribution

In a Bayesian analysis, the “solution/answer” is the posterior distribution on θ. Though, it is relevant to distill down the information it contains. This can be done in the usual ways in which we summarize a probability distribution (similar to the frequentist approach), so by

- point summaries of the
  - central tendency
    - the mean
    - the median
    - the mode
  - variability
    - variance (and standard deviation)
    - range, interquartile range
- intervals or regions as reflection of uncertainty
  - central posterior interval
  - highest posterior density interval / region

Point Summaries/Estimates of Central Tendency

- Mean, $E(\theta|y)$, $\mu_{\theta|y}$
  - Expected a posteriori (EAP) estimator
  - Smallest root mean square error (RMSE) in population defined by $p(\theta)$
- Median, 50th %ile
  - Often preferred in skewed distributions
- Mode, $\hat{\theta} := p(\hat{\theta}|y) = \max p(\theta|y)$
  - Maximum a posteriori (MAP) estimator
  - Somewhat akin to ML, especially with diffuse priors
  - Less used in empirically based estimation (e.g., MCMC)

Point Summaries/Estimates of Variability

- Range, interquartile range
- Posterior variance, $V(\theta|y)$, $\sigma^2_{\theta|y}$; Posterior standard deviation, $\sigma_{\theta|y}$
- Interpretation as the variability of the *parameter*
- Thus while posterior standard deviations may be *numerically similar* to frequentist standard errors, they have *critically different meanings/interpretations*

Posterior intervals

Another common way to summarize the posterior conveying uncertainty is to use intervals of a given posterior probability, say $100(1-\alpha)\%$, this is any interval $[\theta_L, \theta_U]$ such that

$$P(\theta_L \leq \theta \leq \theta_U | y) = 1 - \alpha$$

This is also called a **credibility interval** ("Bayesian confidence interval").

It is somehow the analogue of a confidence interval in classical statistics, but notice the different interpretation, here we say that the unknown parameter lies in the interval with the given probability (rather than saying that the interval is random ...).

(Remind) Interpreting Posterior Credibility Intervals

Posterior credibility intervals, expression of uncertainty, are interpreted as direct probability statements for the unknown parameter.

- The interpretation of the interval is such that probability is *ascribed to the parameter*
- Thus while posterior credibility intervals are often numerically similar to frequentist confidence intervals, they have critically different meanings/interpretations

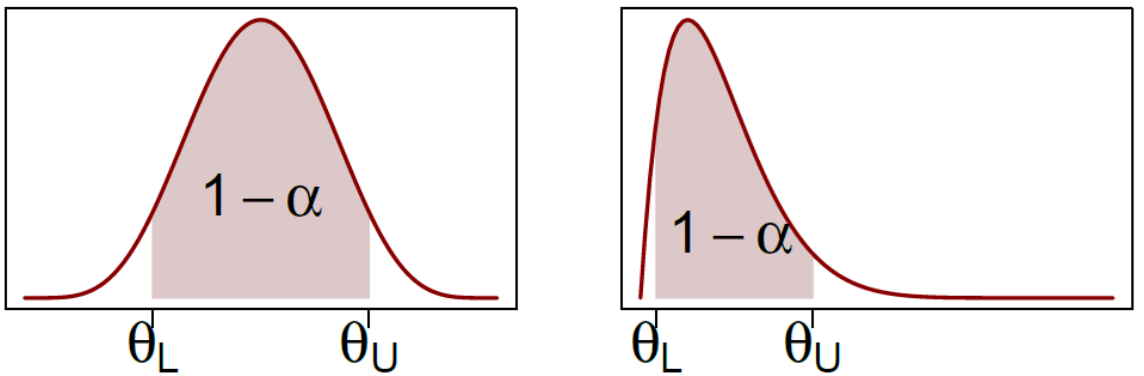
> Adopting an explicitly Bayesian approach would resolve a recurring source of confusion for these researchers, letting them say what they mean and mean what they say.
>
> -- Jackman (2009, p. xxviii)
>
> Frequentist CI theory says nothing at all about the probability that a particular, observed confidence interval contains the true value; it is either 0 (if the interval does not contain the parameter) or 1 (if the interval does contain the true value)…
> …Only the Bayesian procedure…[yielding posterior] credible intervals…allows the interpretation that there is a [X]% probability that the [parameter] is located in the interval.
>
> -- Morey et al. (2016, p. 105, pp. 113-114)

Central Credibility Intervals

A $100(1-\alpha)\%$ **central Interval**, CI, or ***equal-tailed interval*** is the interval of values below and above which the $100\alpha/2\%$ of posterior probability lies

$$I_\alpha = [q_{\alpha/2}, q_{1-\alpha/2}]$$

where q_x is the quantile of order x of $p(\theta|y)$. Below, two examples of central posterior intervals based on quantiles.



Highest Posterior Density Intervals/Regions

Highest Posterior Density, HPD, region: set of values that contains the $100(1-\alpha)$% of posterior probability and for which the density is never lower than the density outside it.

In formulas

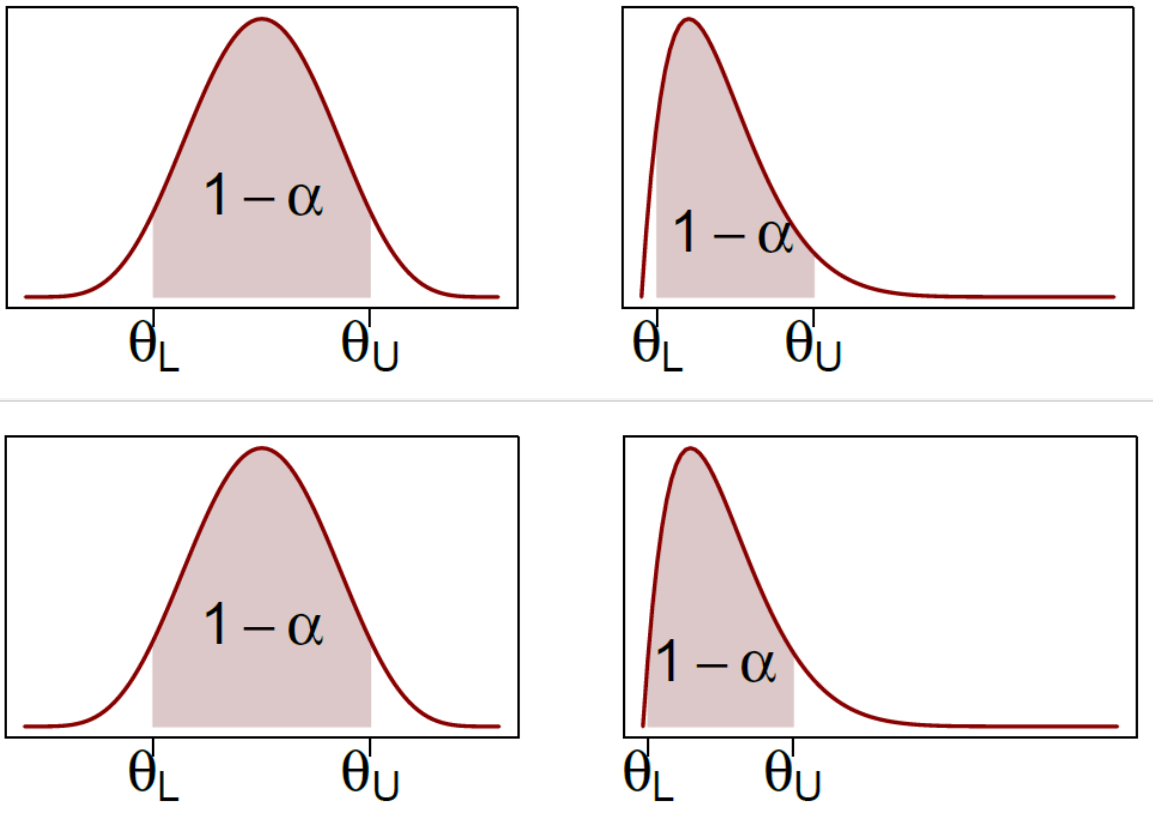
$$\{\theta|\pi(\theta|y) > c_\alpha\}$$

where c_α is such that

$$\int_{\theta|\pi(\theta|y)>c_\alpha} \pi(\theta|y) = 1-\alpha$$

CI $\neq$HPD when posterior is bimodal (/multimodal) or asymmetric
CI $=$HPD when posterior is unimodal and symmetric

Central posterior intervals vs HPD regions



Posterior point estimates for the Uniform-Binomial model

- $E(\theta) = \frac{1}{2}$
- $E(\theta|y) = \frac{y+1}{n+2}$
  The posterior mean represents a compromise between the prior mean, $\frac{1}{2}$, and the observed proportion, $\frac{y}{n}$, and in this compromise data weight increases with their numerosity.
- $\text{Mode}(\theta|y) = \frac{y}{n}$
  The posterior mode is the MLE: since the prior is flat, the maximum of the posterior is where the maximum of the likelihood is.
- $V(\theta|y) = \frac{(y+1)(n-y+1)}{(n+2)^2(n+3)}$
  The posterior variance is less readable, notice that it has n^3 at the denominator and n^2 at the numerator.

Relation between prior and posterior

Note that

- $E(\theta) = E(E(\theta|y))$
- $V(\theta) = E(V(\theta|y)) + V(E(\theta|y))$

  that is, the posterior variance is, **on average**, smaller than the prior variance ($V(\theta) > E(V(\theta|y))$).

  In particular it is smaller the greater is the variation of $E(\theta|y)$ across y.

  Hint to conflicting priors.

This is a general result obtained if we express the mean and variance of a r.v. u in terms of the conditional mean and variance given some related quantity v.

- $E(u) = E(E(u|v))$,
- $V(u) = E(V(u|v)) + V(E(u|v))$

More on posterior summaries for Uniform-Binomial model

We may compute the average over y

$$\begin{aligned} E(V(\theta|y)) &= \frac{1}{(n+2)^2(n+3)} E((y+1)(n-y+1)) \\ &= \frac{1}{(n+2)^2(n+3)} E(ny + n - y^2 + 1) \\ &= \frac{1}{(n+2)^2(n+3)} (n^2/6 + 5n/6 + 1) \end{aligned}$$

(by using the result that the marginal distribution of y is uniform on $(0, n)$.)

Remember that $V(\theta) = \frac{1}{12}$ and if $n = 1$, $E(V(\theta|y)) = \frac{1}{18}$

Prediction for a future Bernoulli trial

Consider a new observation $\tilde{y}$, which behaves like the y_i, that is

- $\tilde{y}$ is independent of $y_1, \ldots, y_n$ conditional on θ
- $P(\tilde{y} = 1|\theta) = \theta$

then the prior predictive distribution is

$$P(\tilde{y} = 1) = \int_0^1 \theta p(\theta) d\theta = \int_0^1 \theta d\theta = E(\theta) = 1/2$$

while the posterior predictive distribution is

$$P(\tilde{y} = 1|y) = \int_0^1 \theta p(\theta|y) * d\theta = E(\theta|y) = \frac{y+1}{n+2}$$

Extreme cases

$$p(\tilde{y} = 1|y = 0) = \frac{1}{n+2} \quad \text{and} \quad p(\tilde{y} = 1|y = n) = \frac{n+1}{n+2}$$

- cf. maximum likelihood

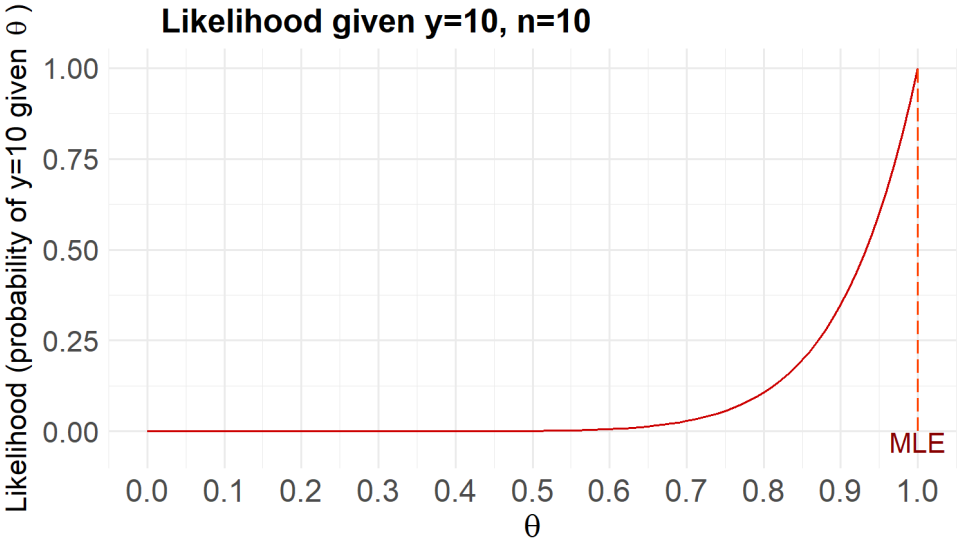
Benefici dell'integrazione

Consider the number of correct responses in the set of responses to the J equally difficult tasks.

Example: *Perfect Response Patterns* ($n = 10, y = 10$)

- An examinee correctly completes all 10 tasks
- What should you believe about the examinee's proclivity to complete tasks?
- Likelihood vs Bayes with minimal prior information

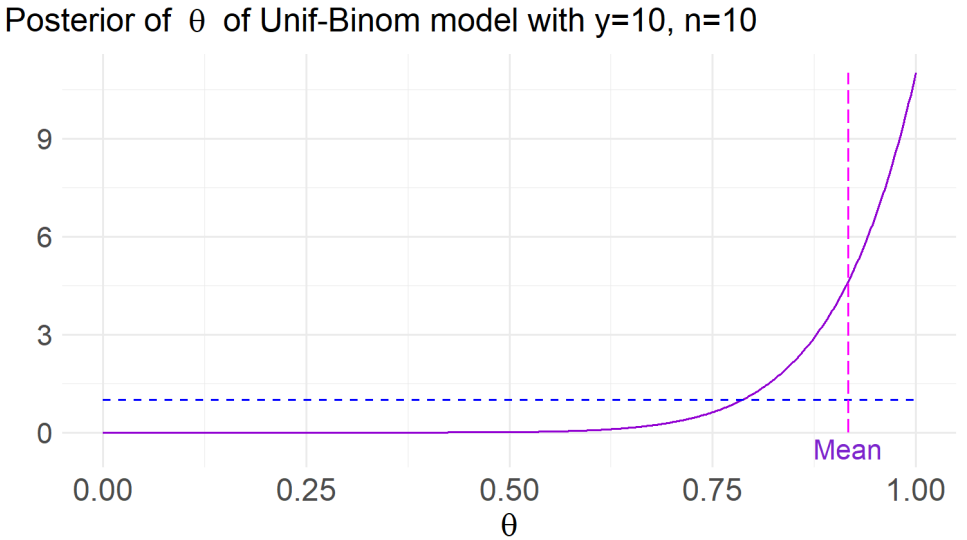
Perfect Response Pattern: ML



- The maximum likelihood estimate (MLE) is 1.0
- This is a boundary: problems arise with standard errors, sampling distribution, hypothesis testing
- Do we really think this is a good estimate of an examinee's proclivity to correctly complete tasks? Do we really think the examinee will correctly complete every single task?

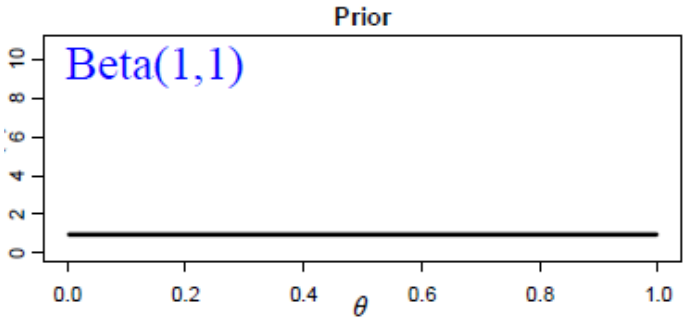
Perfect Response Pattern: Bayes with minimal prior information

- Uniform prior: $U(0, 1)$
- Beta posterior: $Beta(11, 1)$

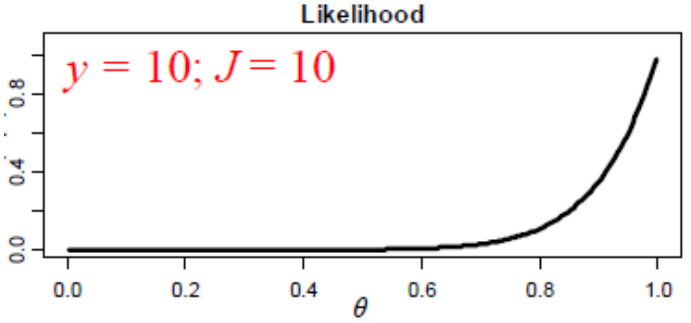


- $E(\theta|y) = 11/12$

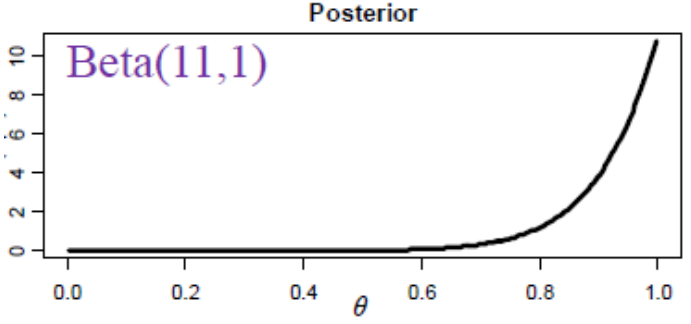
Example: ***Perfect Response Patterns***



Prior mean = .5
Prior mode = NA
Prior sd = .29



Max. likelihood = 1



Posterior mean = .92
Posterior mode = 1
Posterior sd = .08

Note: Prior predictive distribution for y

With a uniform prior on θ, the distribution of a Bernoulli $\tilde{y}$ prior to observing the data is $P(\tilde{y} = 1) = \frac{1}{2}$.

For a binomial $y = (\sum_{i=1}^{n} y_i) \sim \text{Bin}(n, \theta)$

$$\begin{aligned} p(y) &= \int_0^1 p(y|\theta)p(\theta)d\theta \\ &= \binom{n}{y} \int_0^1 \theta^y (1-\theta)^{n-y} p(\theta) d\theta \\ &= \binom{n}{y} \frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)} = \\ &= \frac{1}{n+1} \end{aligned}$$

Justification of the uniform prior

$p(\theta)$=1 if

- we want a uniform prior predictive distribution; in the binomial example the Bayesian reasoning entails:

$$p(y) = \frac{1}{1+n} \quad y = 0, 1, \ldots, n$$

  - justification based on the observables y and n
  - justification of Bayes

- we think that all values of θ are equally probable; *principle of insufficient reason* (Laplace), i.e. "If nothing is known about θ then the uniform is appropriate".

  - justification based on the unobservable θ

Prior

We considered a uniform prior on θ, this has been the choice of both Bayes and Laplace, who (loosely speaking) justified it

- Bayes based on the fact that it implies a uniform predictive prior on y
- Laplace based on the so called 'principle of insufficient reason' because he had no information about θ

Afterwards different approaches to the prior specification have been considered, in what follows we discuss different choices and look at their consequences, keeping in mind the following

- a prior need only to reasonably summarize the knowledge we have on θ
- if this information is scarce, the effect of the prior should vanish as enough data are collected
- sometimes we have relevant prior information, and we want to use it, sometimes we want to use a prior that is weakly informative, sometimes we want to use a prior that is non-informative, or sometimes we want to use a prior that is computationally convenient

Conjugate priors

A convenient type of prior is the kind that leads to a posterior in the same family, this property is called **conjugacy**.

- This is not available for any likelihood (just for exponential distributions, plus some irregular cases),

Used for computational reasons, and still sometimes used for special models to allow partial analytical marginalization

Definition

If $\mathcal{F}$ is the class of sampling distributions and $\mathcal{P}$ is the class of prior distributions, $\mathcal{P}$ is a **natural conjugate** for $\mathcal{F}$ if $\mathcal{P}$ is the set of all densities having the same functional form in θ as the likelihood.

Conjugate priors are useful because

- it is easy to obtain the results (analytic forms for the mean, variance, etc.)
- they simplify the calculations
- they are a good starting point
- you can use mixtures of conjugate families

Conjugate priors and exponential families

Probability distributions belonging to an exponential family have natural conjugate prior distributions.

Definition

A family of distributions $\mathcal{F} = \{p(y|\theta) : \theta \in \Theta \subset \mathbb{R}^d\}$ is an exponential family if all its members have the form

$$p(y|\theta) = f(y)g(\theta)\exp^{\phi(\theta)^T u(y)}$$

where $f : \mathbb{R} \to \mathbb{R}$, $g : \mathbb{R}^d \to \mathbb{R}$, $\phi : \mathbb{R}^d \to \mathbb{R}^d$ and $u : \mathbb{R}^d \to \mathbb{R}^d$ are known functions.

$\phi(\theta)$ is called the natural parameter of $\mathcal{F}$.

Exponential family: likelihood and sufficient statistic

If a vector of observations $y = (y_1, \ldots, y_n)$ is observed and y_i are iid following a distribution from $\mathcal{F}$

$$p(y|\theta) = \left(\prod_{i=1}^{n} f(y_i) \right) g(\theta)^n \exp\left(\phi(\theta)^T \sum_{i=1}^{n} u(y_i) \right)$$

hence

$$p(y|\theta) \propto g(\theta)^n \exp\big(\phi(\theta)^T t(y) \big)$$

where

$$t(y) = \sum_{i=1}^{n} u(y_i)$$

is a **sufficient statistic**.

The quantity $t(y)$ is called a sufficient statistic for θ, because the likelihood for θ depends on the data y only through the value of $t(y)$.

Conjugate distribution for an exponential family

If the prior is of the form

$$p(\theta) \propto g^{\eta}(\theta) \exp(\phi(\theta)^T \nu)$$

then the posterior is

$$p(\theta|y) \propto g^{n+\eta}(\theta) \exp(\phi(\theta)^T (t(y) + \nu))$$

which has the same form as the prior.

It can be shown that only exponential families of distributions have natural conjugate priors.

(That is because have a fixed number of sufficient statistics)

Beta-Binomial model

The conjugate prior for the Binomial model is the Beta distribution:

If $\theta \sim \text{Beta}(\alpha, \beta)$ then $\theta|y \sim \text{Beta}(\alpha + y, \beta + n - y)$

as is easily checked:

$$\begin{aligned} p(\theta|y) &\propto \theta^y (1-\theta)^{n-y} \theta^{\alpha-1} (1-\theta)^{\beta-1} \\ &= \theta^{y+\alpha-1} (1-\theta)^{n-y+\beta-1} \end{aligned}$$

- $(\alpha - 1)$ e $(\beta - 1)$ can be interpreted as the number of *prior successes* and *prior failures*, and $\alpha + \beta - 2$ as the number of *prior observations*
- Uniform prior when $\alpha = 1$ and $\beta = 1$

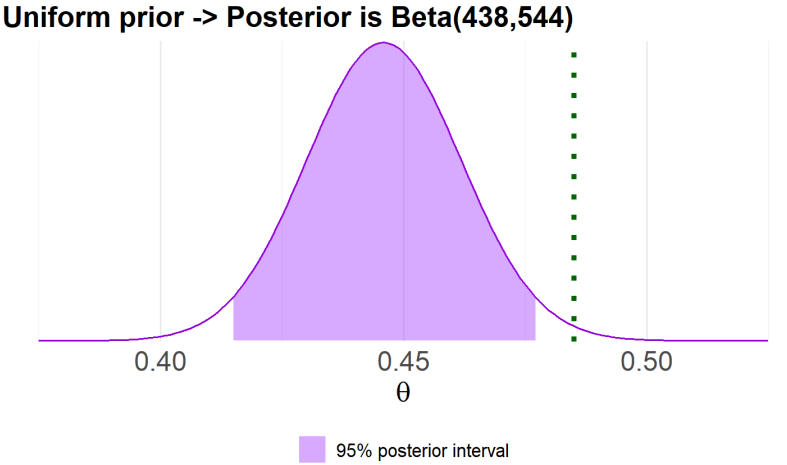
Example: placenta previa

Probability of a female birth with placenta previa (BDA3 p. 37)

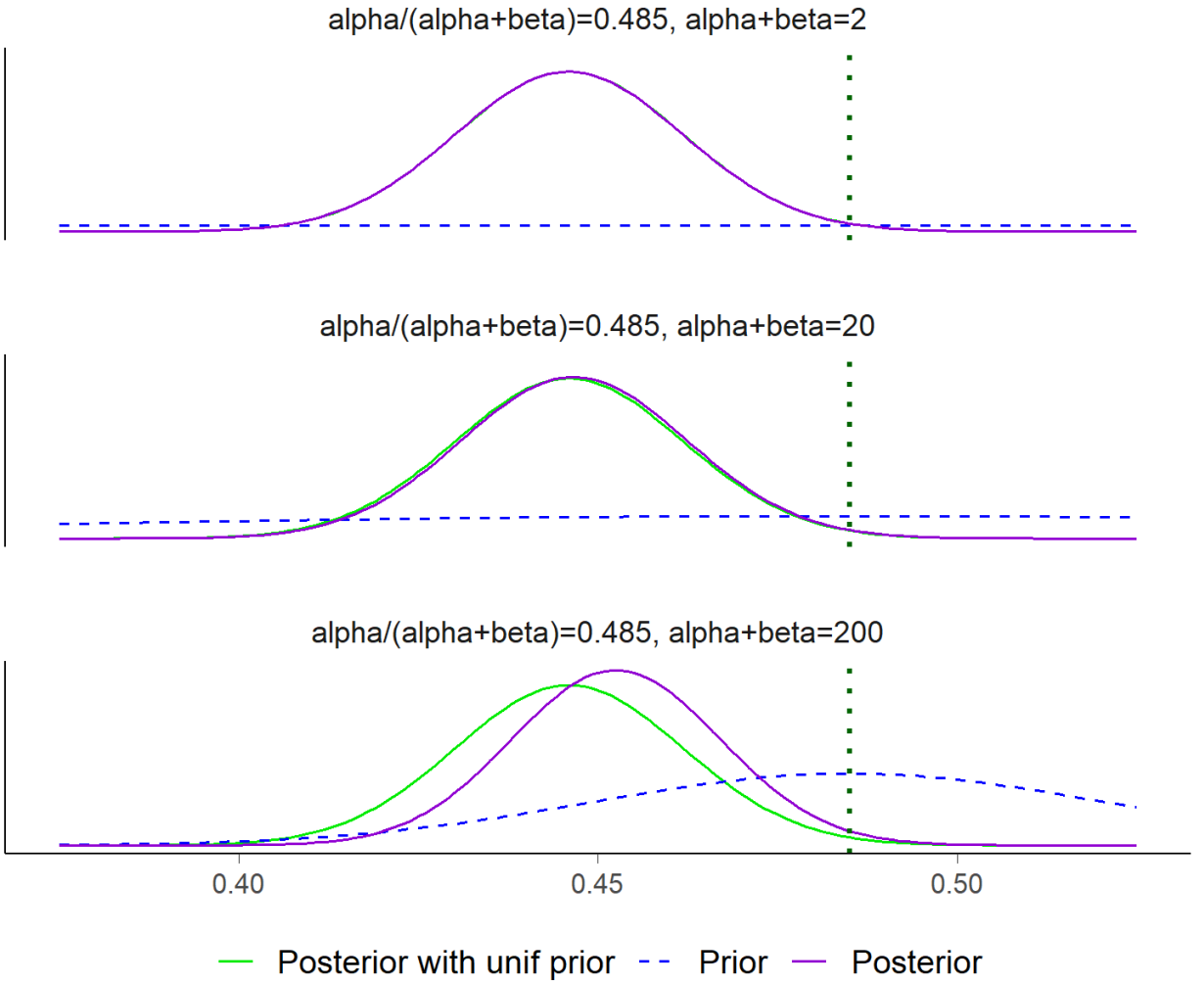
In a study in Germany, 437 females out of 980 births with placenta previa were observed.

Do we have evidence that the proportion of female births with placenta previa is less than 0.485, which is the corresponding proportion in the general population?

- (empirical proportion is 0.445)
- Likelihood: $\propto \theta^{437}(1-\theta)^{980-437}$
- Prior: $U(\theta|0,1) = Beta(\theta|1,1)$
- Posterior: $\propto \theta^{437}(1-\theta)^{980-437}$, $Beta(\theta|438,544)$
- $P(\theta < 0.485|y) = 0.9928$

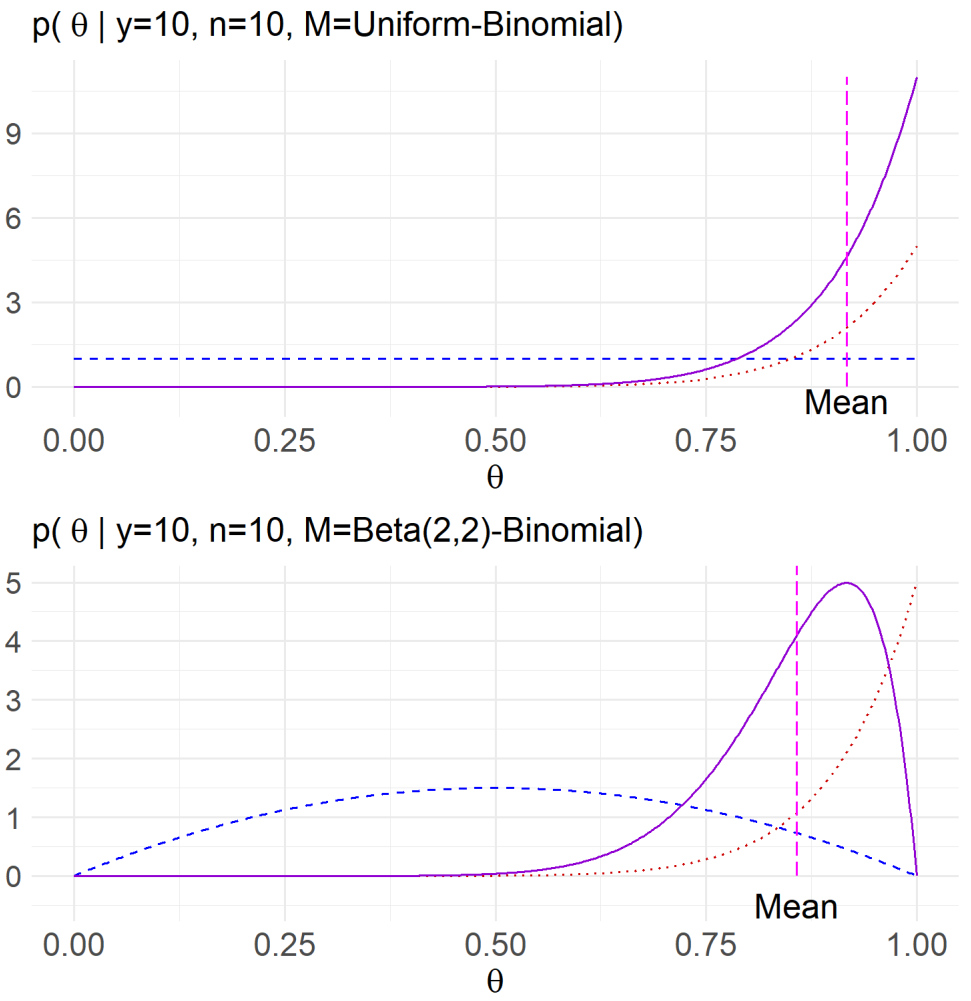


- Comparison of posterior distributions with different parameter values for the Beta prior distribution: Beta priors centered in the population mean, 0.485, and with increasing strenght



Still on Perfect Response Pattern example

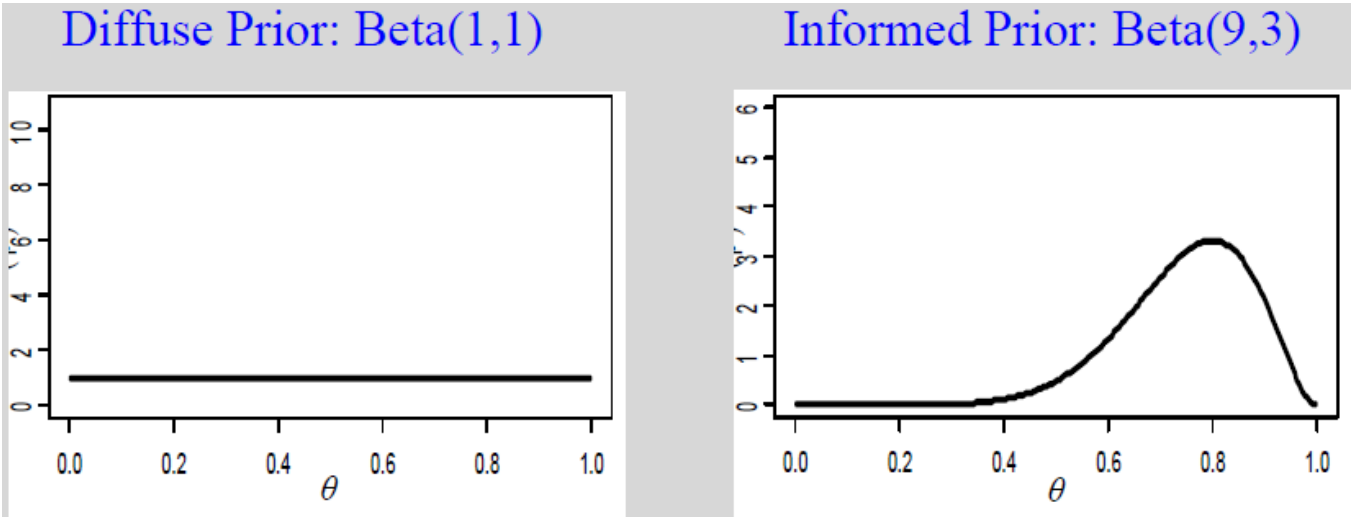
Example: $n = 10, y = 10$ - uniform priori vs Beta(2,2)



Esempio: ***Perfect Response Patterns***

Bayes with an informative prior

- Do you believe a priori that the candidate is very capable of successfully completing these tasks?
- What should you believe about the candidate's ability to complete the tasks?

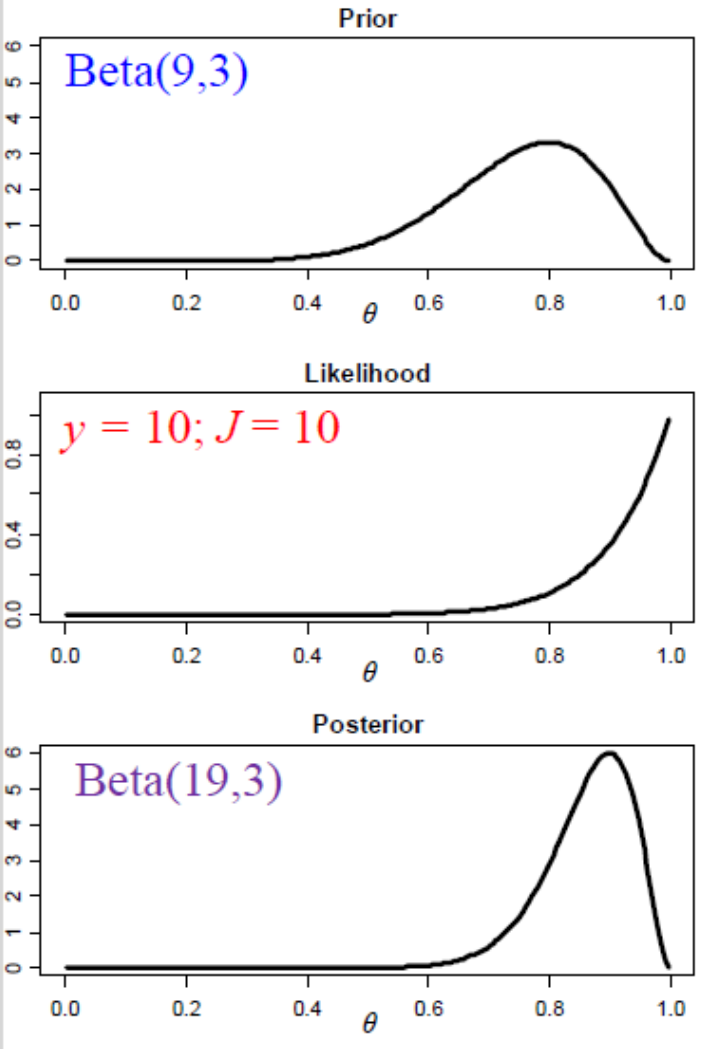


Esempio: ***Perfect Response Patterns***

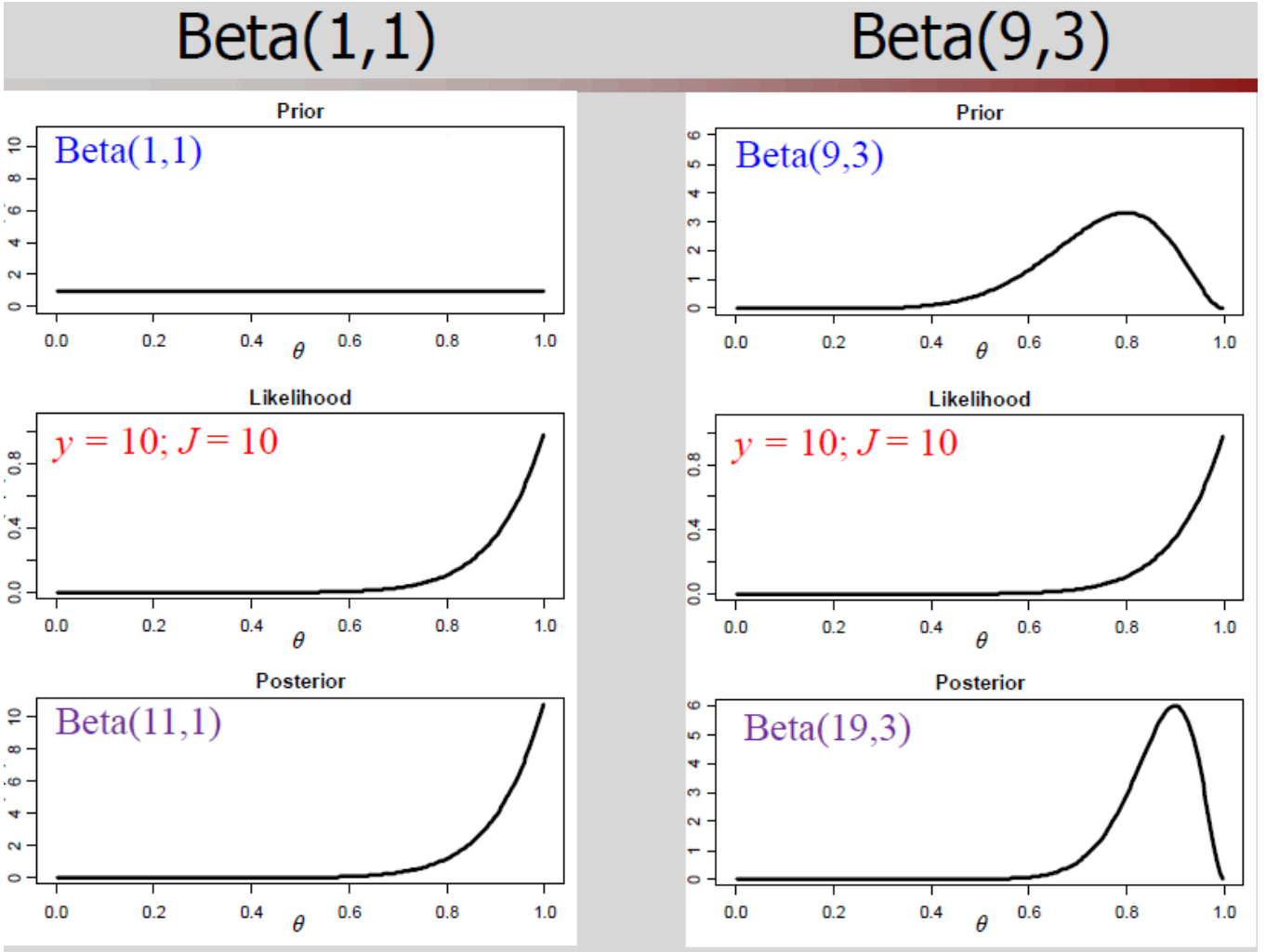
Prior mean = .75
Prior mode = .80
Prior sd = .12

Max. likelihood = 1

Posterior mean = .86
Posterior mode = .90
Posterior sd = .07



Esempio: ***Perfect Response Patterns***



Beta Binomial model: Posterior mean

Let us synthesize the posterior distribution using the expectation

$$\begin{aligned} E(\theta|y) &= \int \theta\pi(\theta|y)d(\theta) = \frac{\alpha+y}{\alpha+\beta+n} \\ &= \frac{\alpha+\beta}{\alpha+\beta+n}\frac{\alpha}{\alpha+\beta} + \frac{n}{\alpha+\beta+n}\frac{y}{n} \\ &= \frac{\alpha+\beta}{\alpha+\beta+n}\underbrace{E(\theta)}_{\text{prior mean}} + \frac{n}{\alpha+\beta+n}\underbrace{\frac{y}{n}}_{MLE} \end{aligned}$$

The posterior mean is a weighted average of the prior expectation and the ML estimate, where

- ML estimate prevails if n is large;
- ML estimate prevails if α and β are small:
  - the variance of the prior distribution is large
  - $\alpha+\beta(-2)$, the equivalent number of observation of the prior distribution, is small.
- if $n \to \infty$, $\mathrm{E}[\theta|y] \to y/n$

Beta Binomial model: Posterior variance

The posterior variance is

$$V(\theta|y) = \frac{(\alpha + y)(\beta + n - y)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)} = \frac{E(\theta|y)(1 - E(\theta|y))}{\alpha + \beta + n + 1}$$

- decreases as n increases
- if $n \to \infty$, $\text{Var}[\theta|y] \to 0$
- As y and n gets big
  - $E(\theta|y) \approx y/n$
  - $V(\theta|y) \approx \frac{1}{n}\frac{y}{n}\left(1 - \frac{y}{n}\right)$

Distribuzioni a posteriori quando la dimensione compionaria $\to \infty$

$$p(\theta|y) = \text{Beta}(\theta|\alpha + y, \beta + n - y)$$

$$\overset{n \to \infty}{p(\theta|y) \propto p(y|\theta)p(\theta) \implies p(\theta|y) \propto p(y|\theta)}$$

All'aumentare della dimensione del campione, la distribuzione a posteriori diventa sempre più simile alla verosimiglianza [di solito]

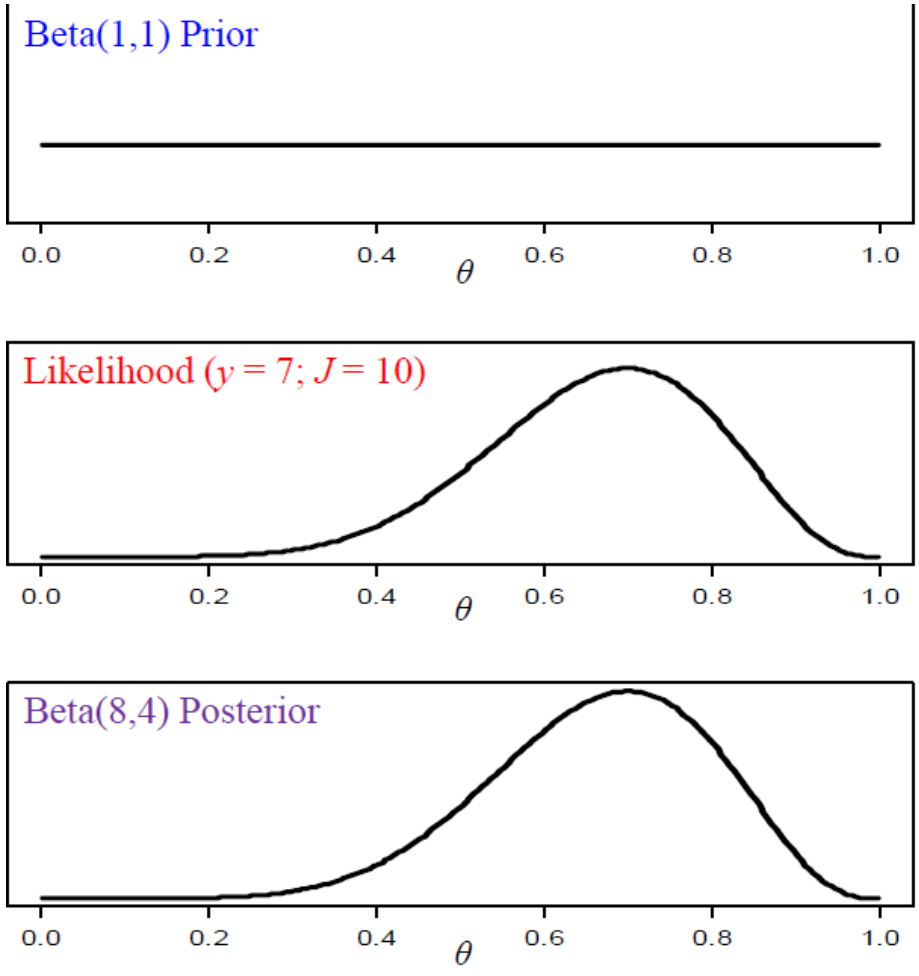
Esempio ($n = 10, y = 7$)

Osserviamo il numero di successi nelle n prove

- Ad esempio, 7 risposte corrette su 10 compiti
- Iniziamo con una a priori uniforme $Beta(1, 1)$ (come dire: 0 successi a priori e 0 fallimenti a priori) e via via costruiamo a priori più informative
- Calcoliamo la a posteriori

Esempio ($n = 10, y = 7$)

A priori uniforme



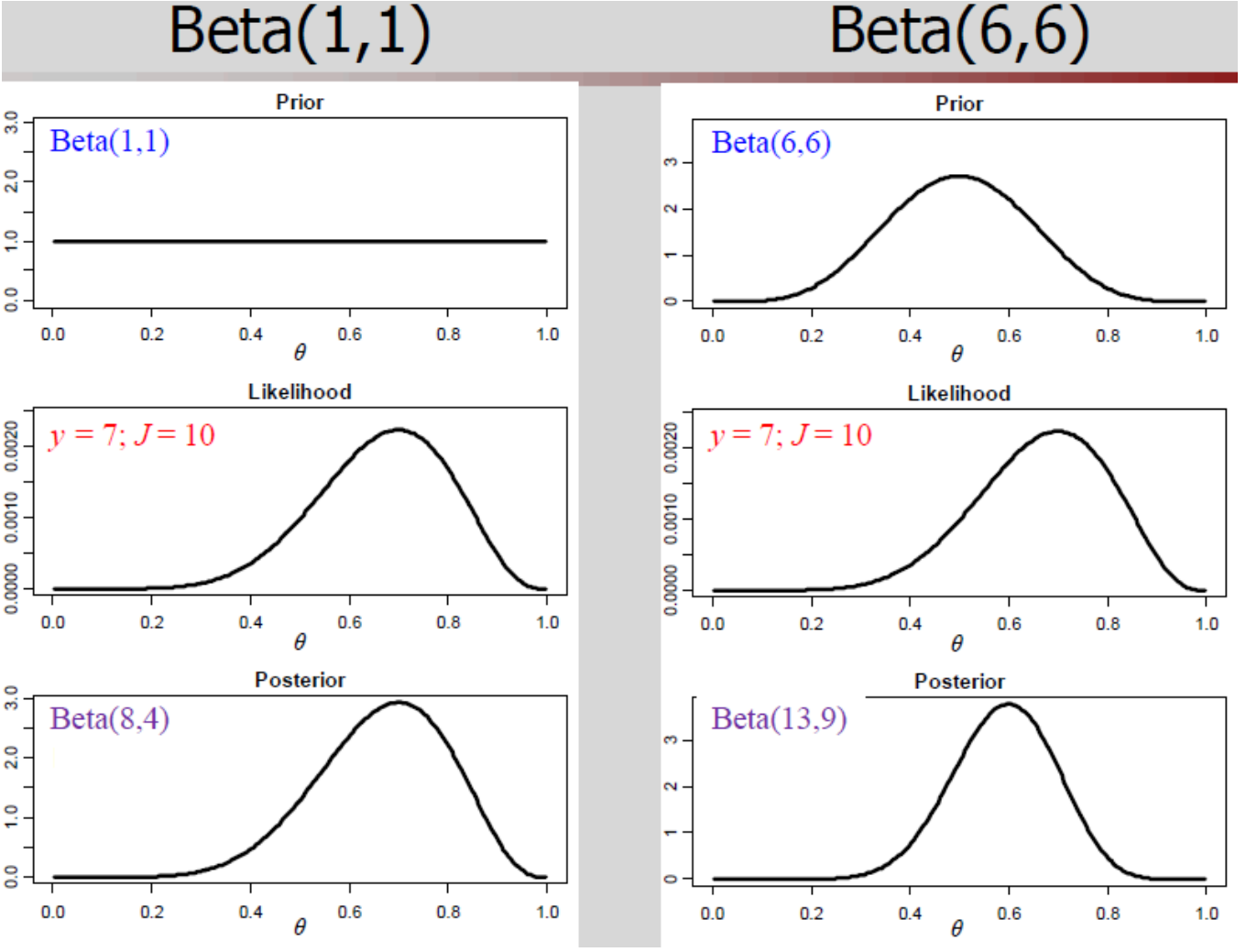
Prior Mean
= 1/2 = .50

Mean of the Data
= 7/10 = .70

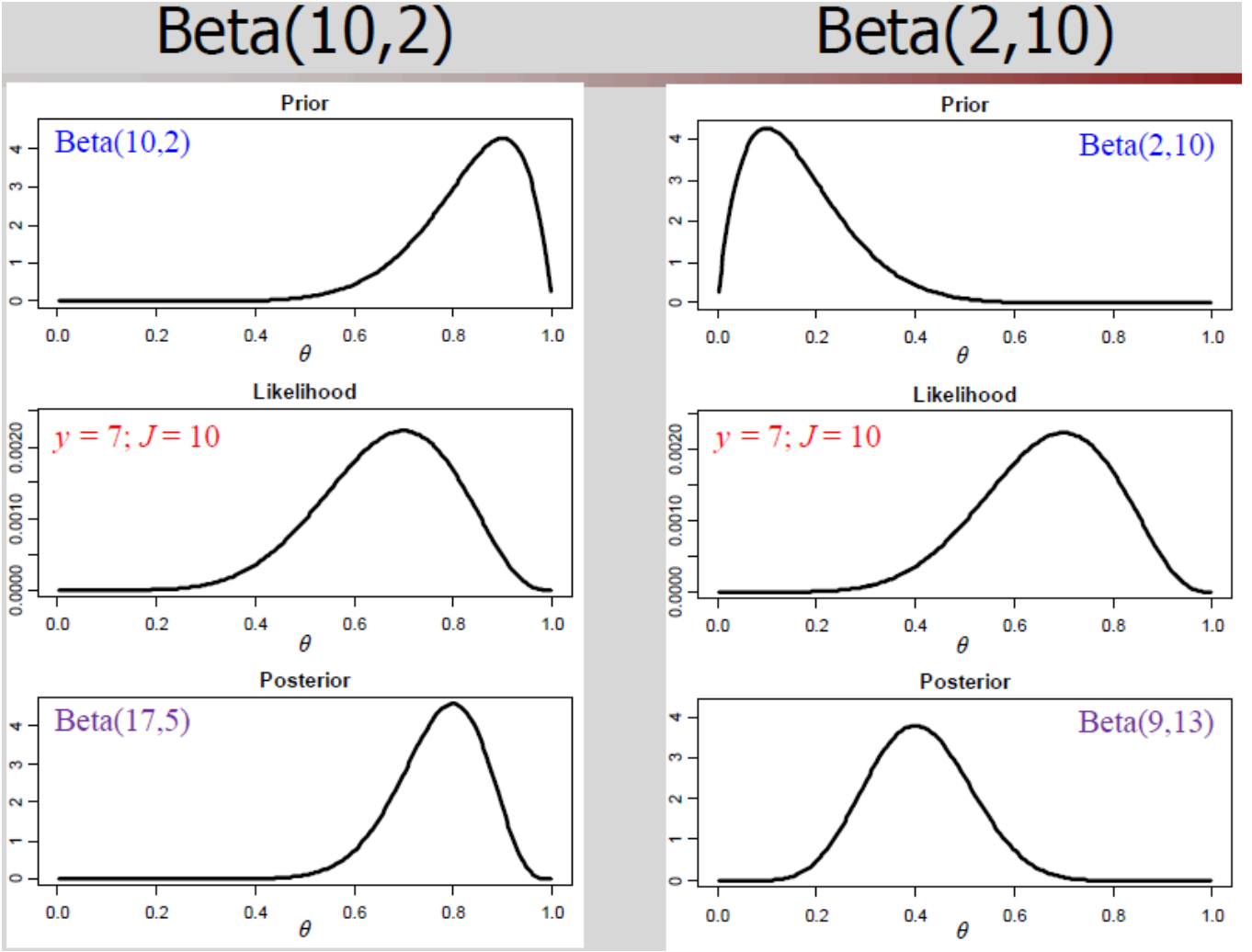
Posterior Mean
= 8/12 = .67

Posterior as synthesis of prior and likelihood

Effetti dell'informazione nella a priori

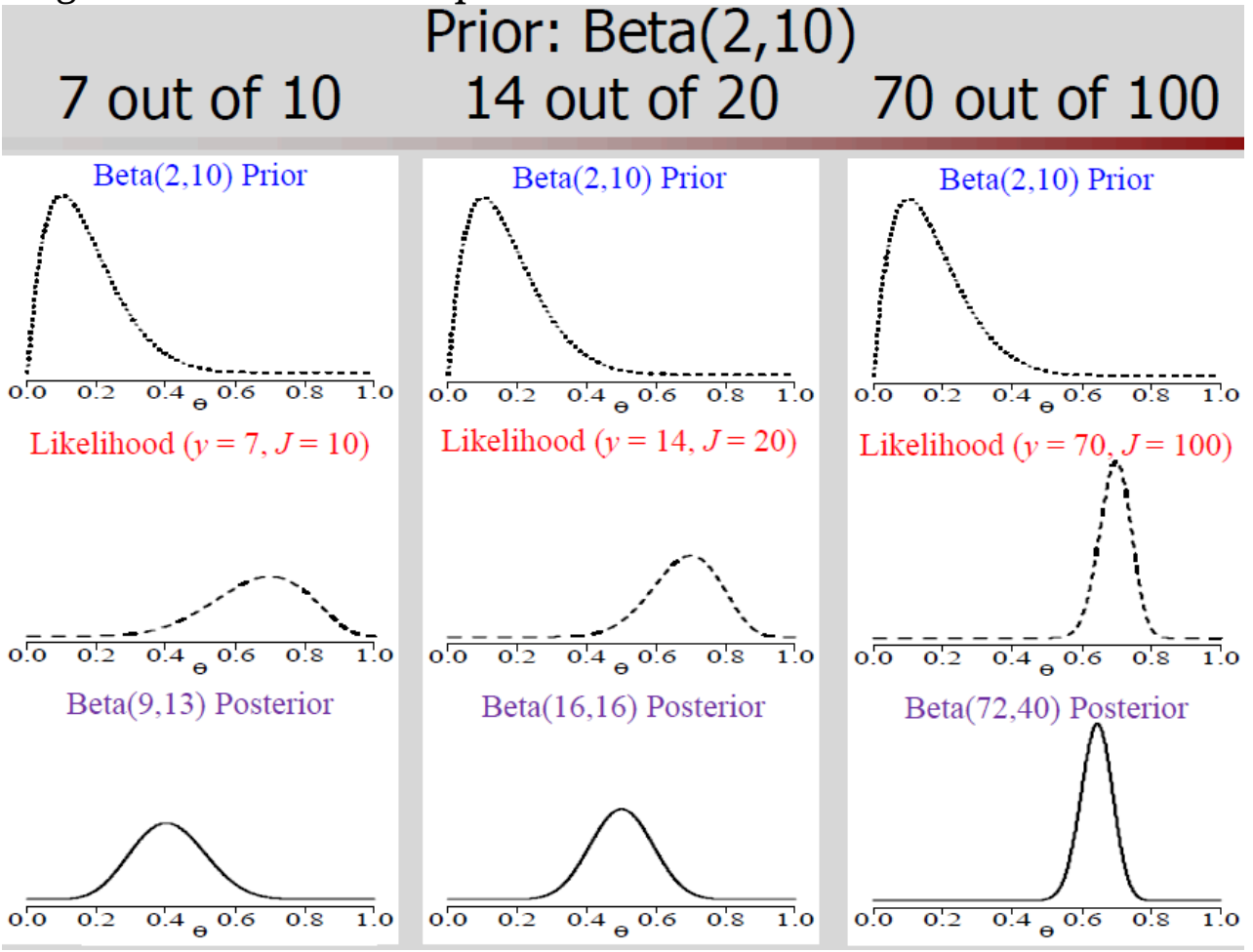


Effetti dell'informazione nella a priori



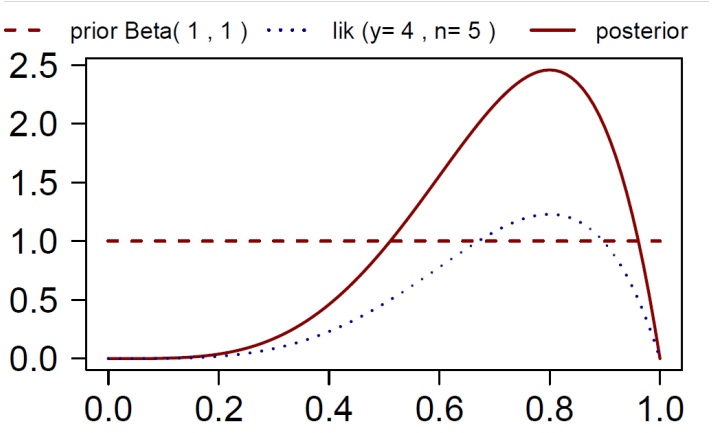
Effetti dell'informazione nei dati

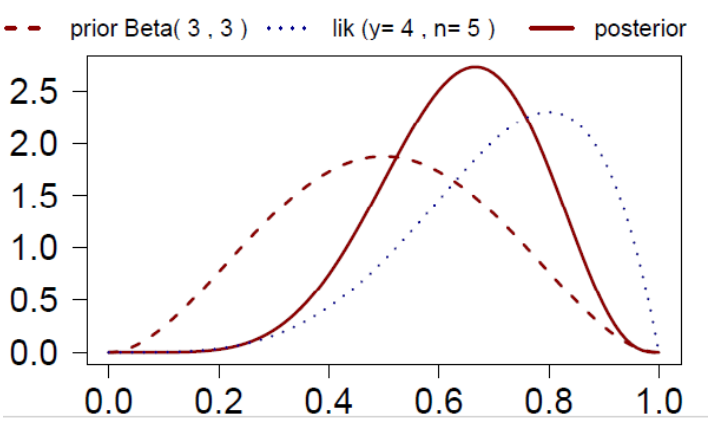
Mantenendo lo stesso tasso di successo nel campione (cioè, y/n è costante), all'aumentare di n, α e β diminuiscono di importanza e i dati / la verosimiglianza dominano l'a posteriori.

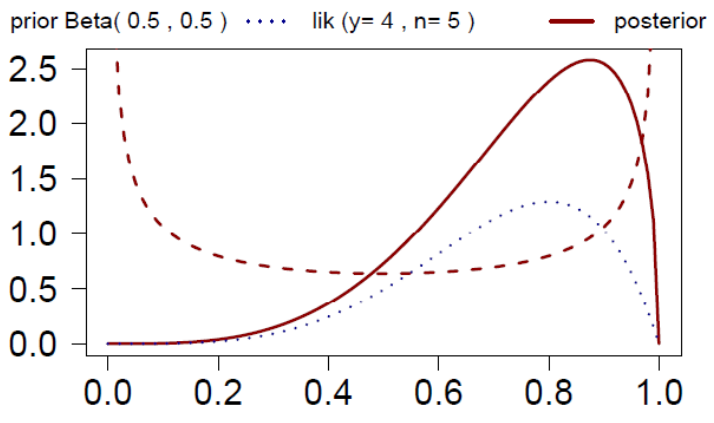


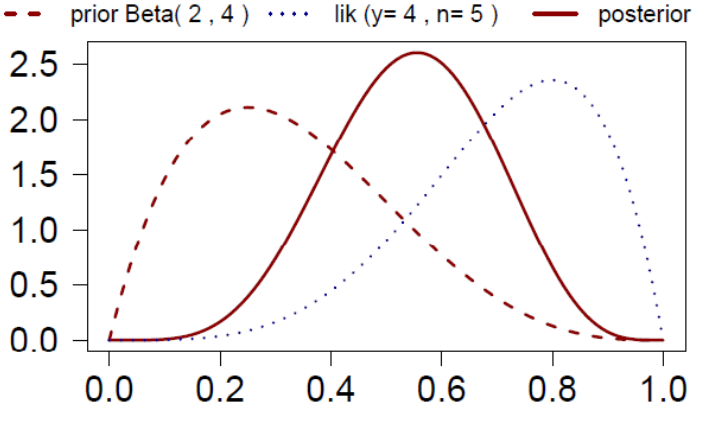
Different priors ⇒ different posteriors

uniform, centered at 0.5, 0-1, rightly asymmetric





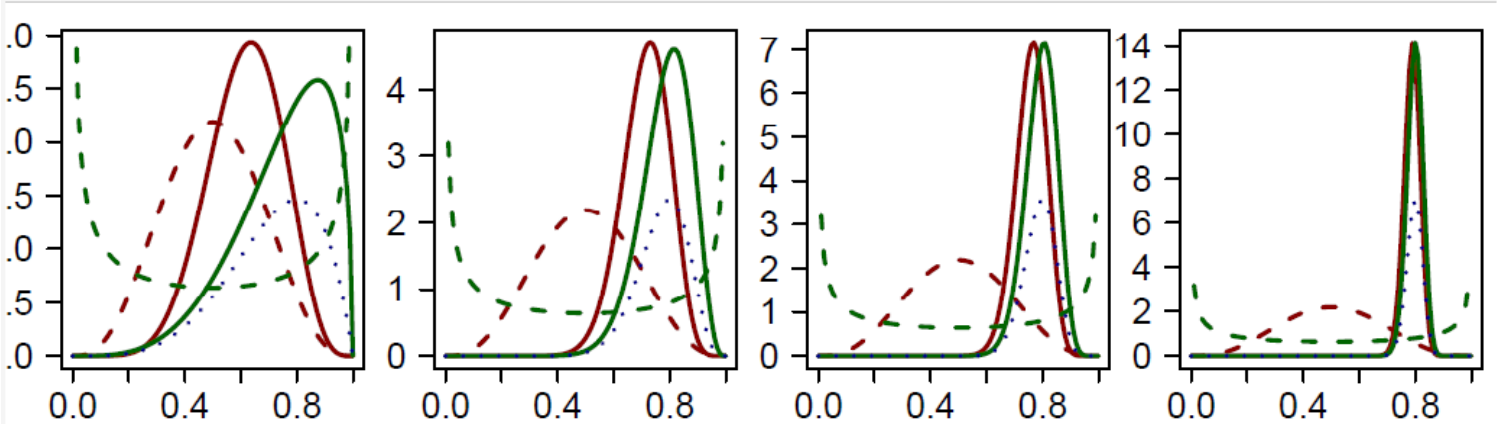




Prior effect as n increases

The effect of the prior, however, tend to disappear as enough sample information is entered.

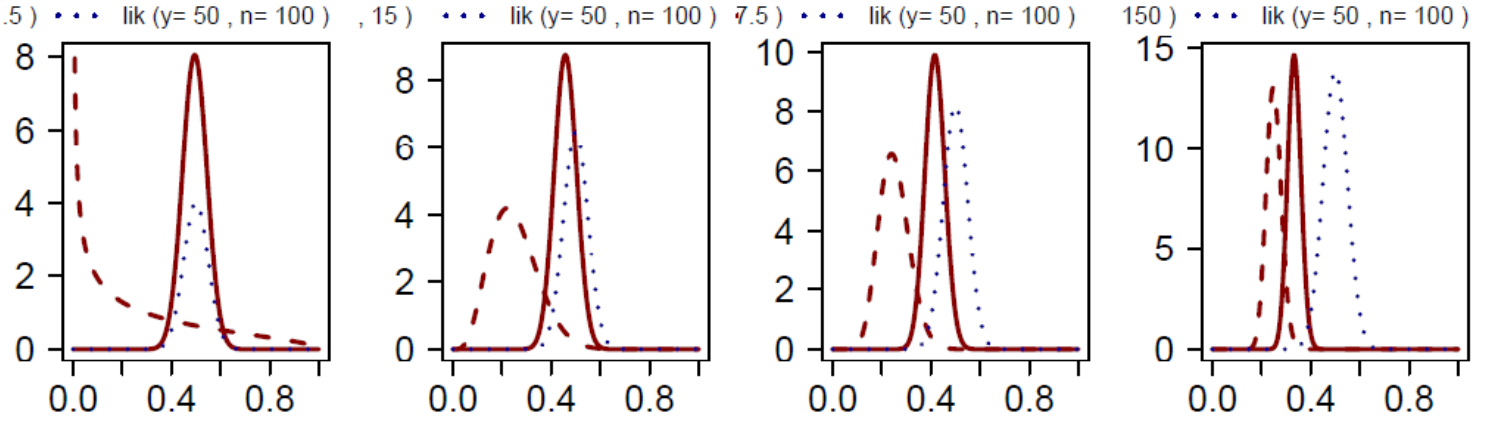
In the following we observe the effect on the posterior of two distinct priors on samples of $n = 5, 20, 50, 200$, always with $y/n = 0.8$



Prior effect as $\alpha + \beta$ increases

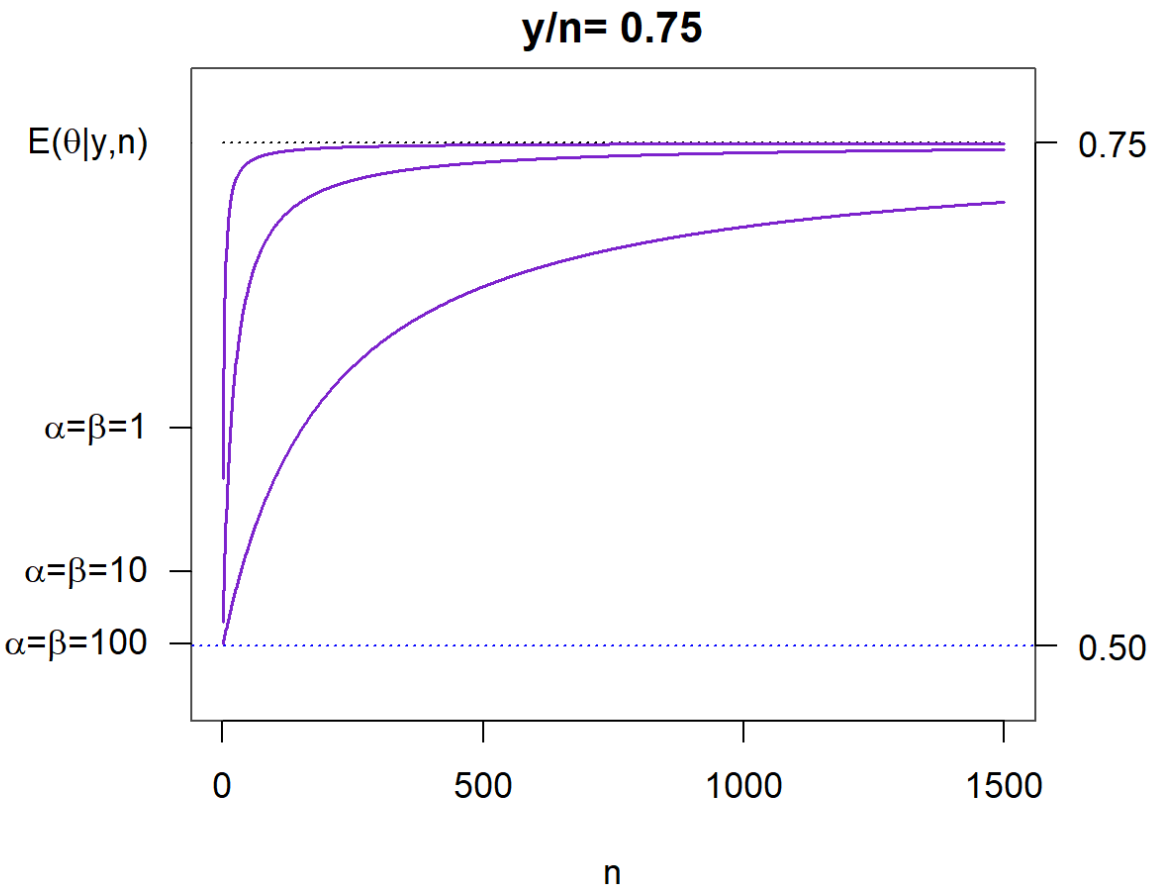
We can see things from another point of view and consider different priors with the same sample.

We observe a sample with $n = 100$ and $y = 50$, the prior mean is 0.25, $\alpha + \beta$ is 2, 20, 50, 200



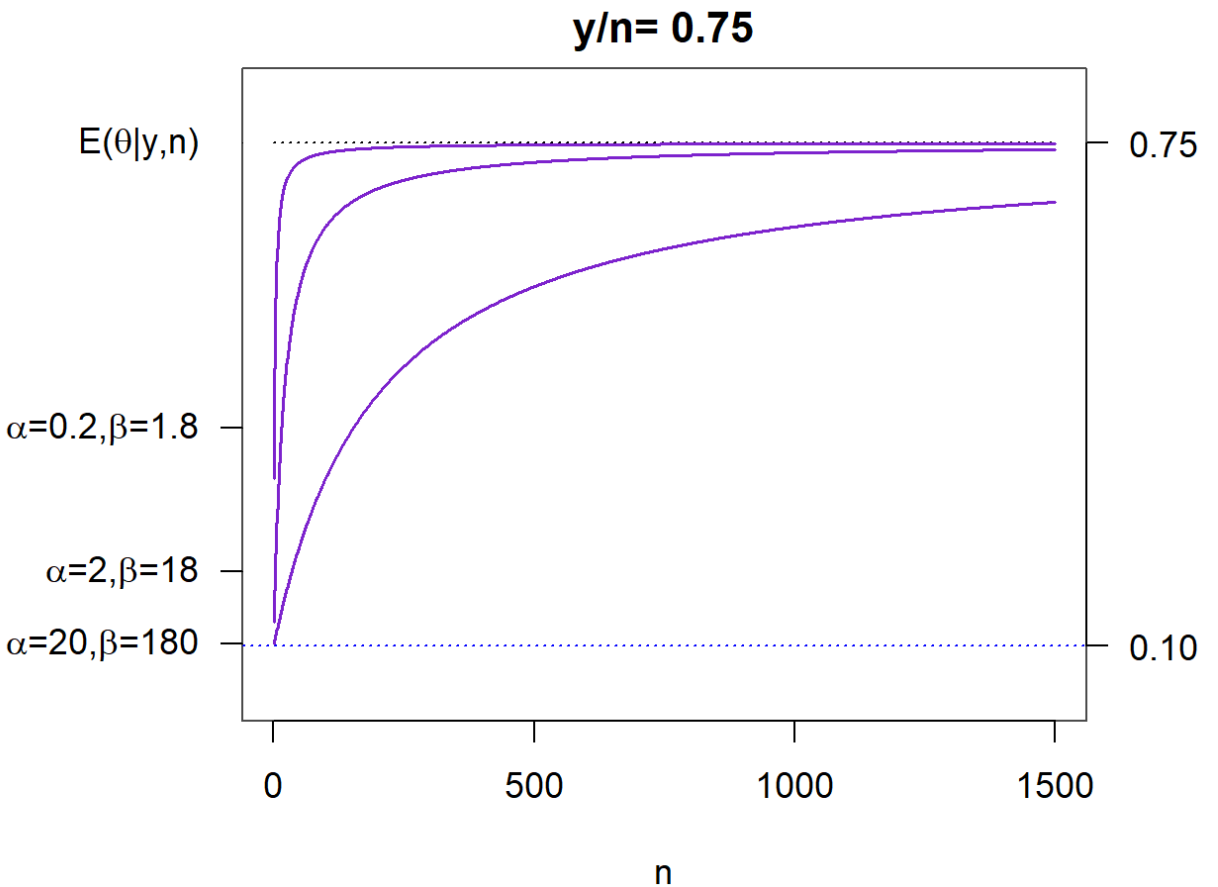
Posterior mean as a function of sample size

Posterior mean as n increases for different priors



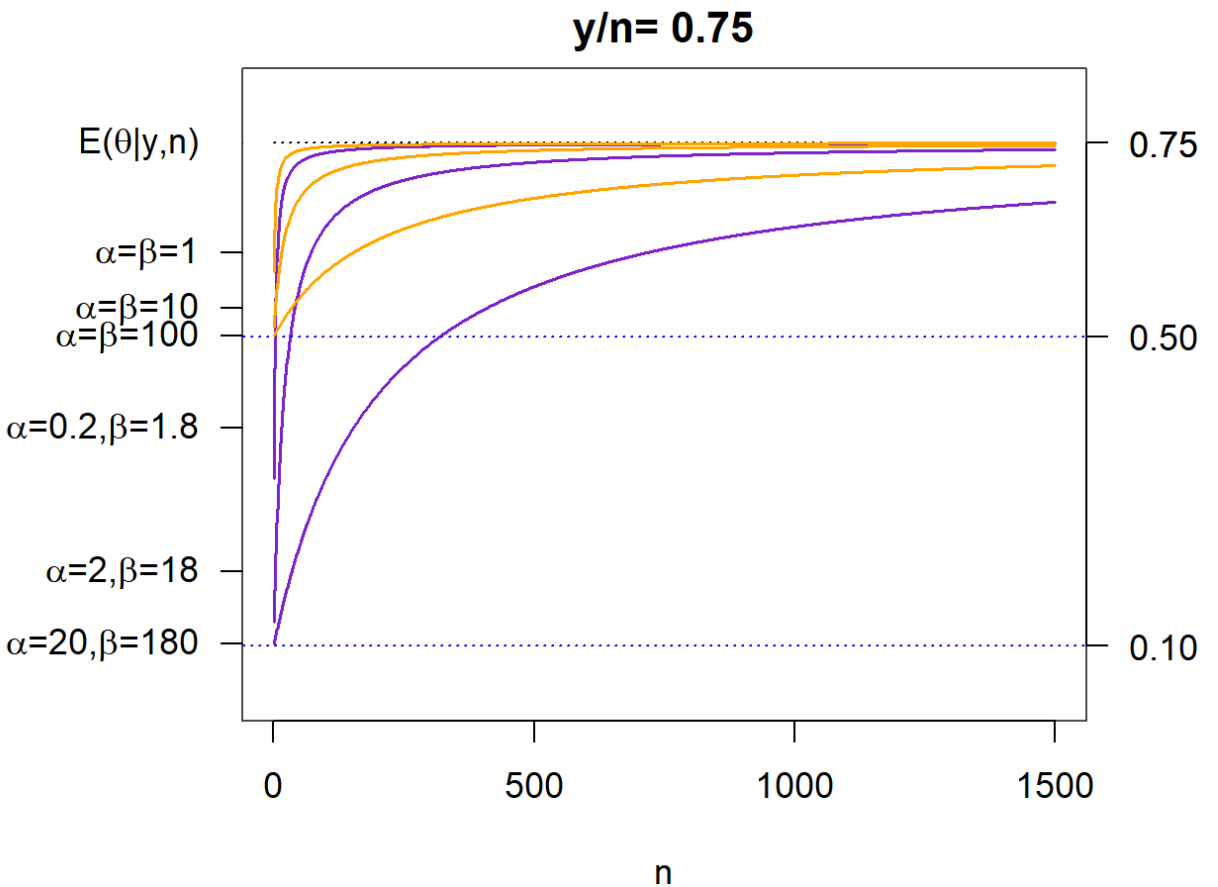
Posterior mean: conflicting priors

Posterior mean as n increases for different priors



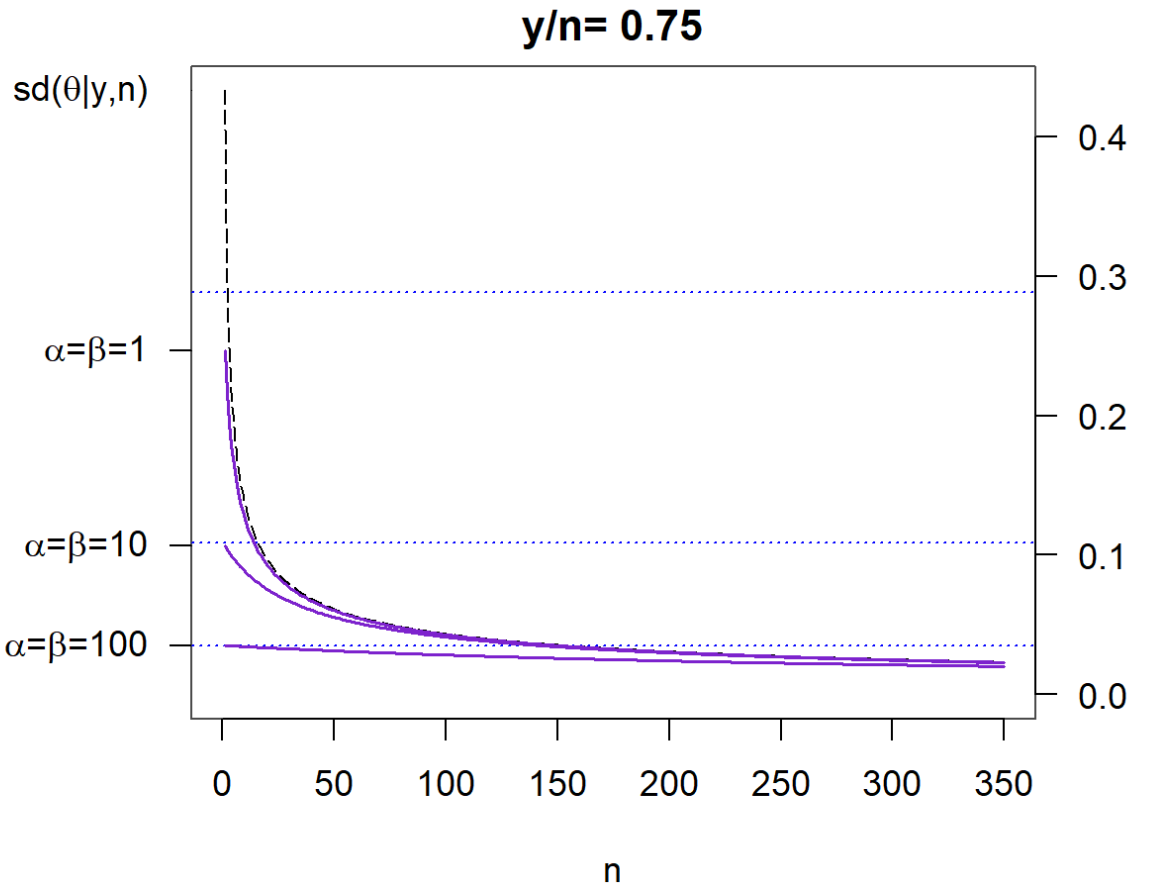
Posterior mean: all together

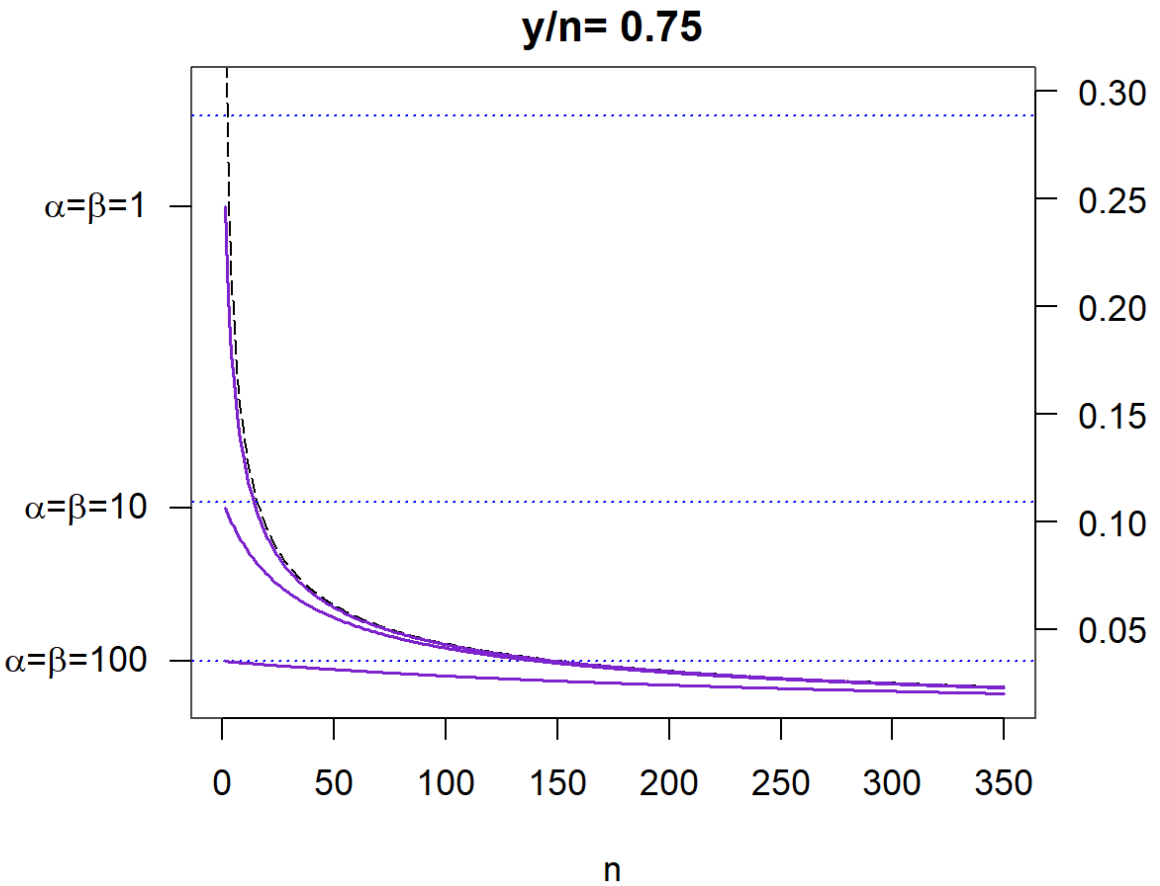
Posterior mean as n increases for different priors



Posterior variance as a function of sample size

Posterior variance as n increases for different priors

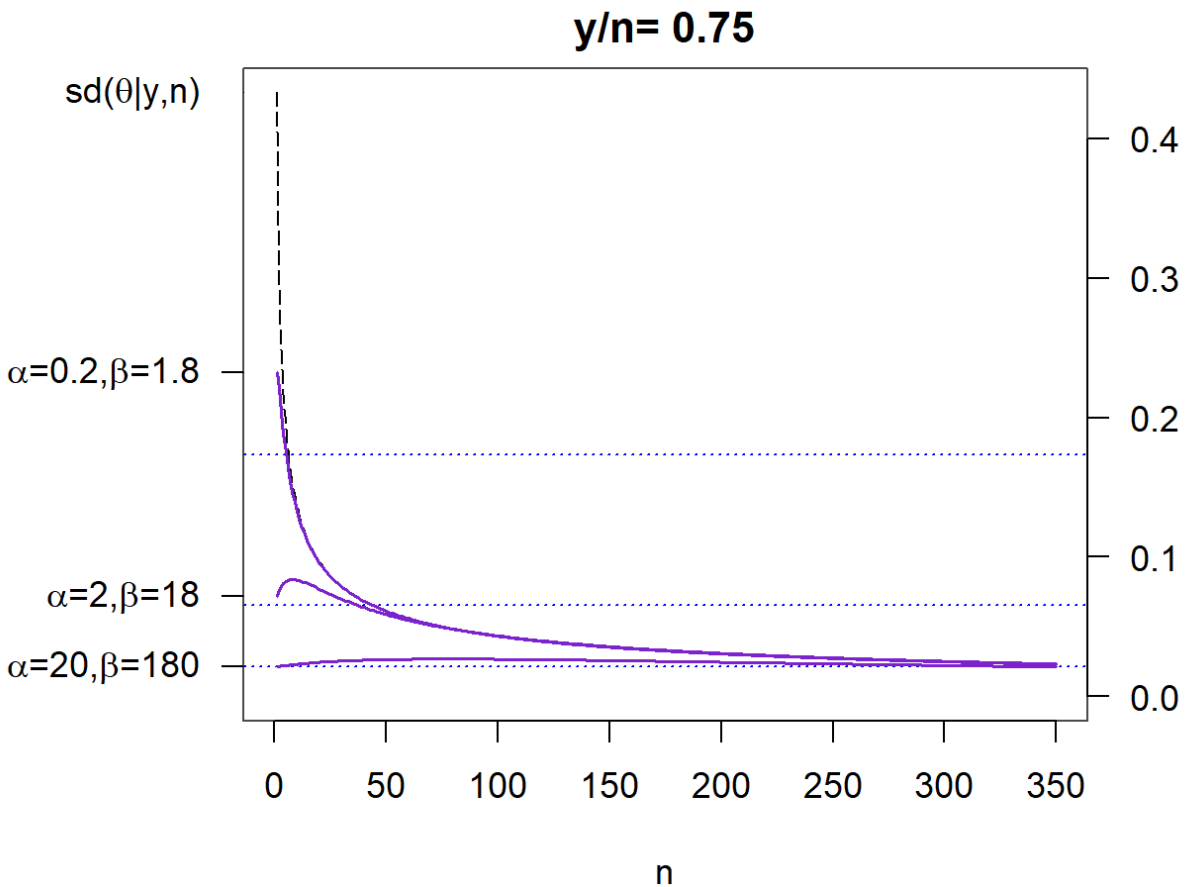


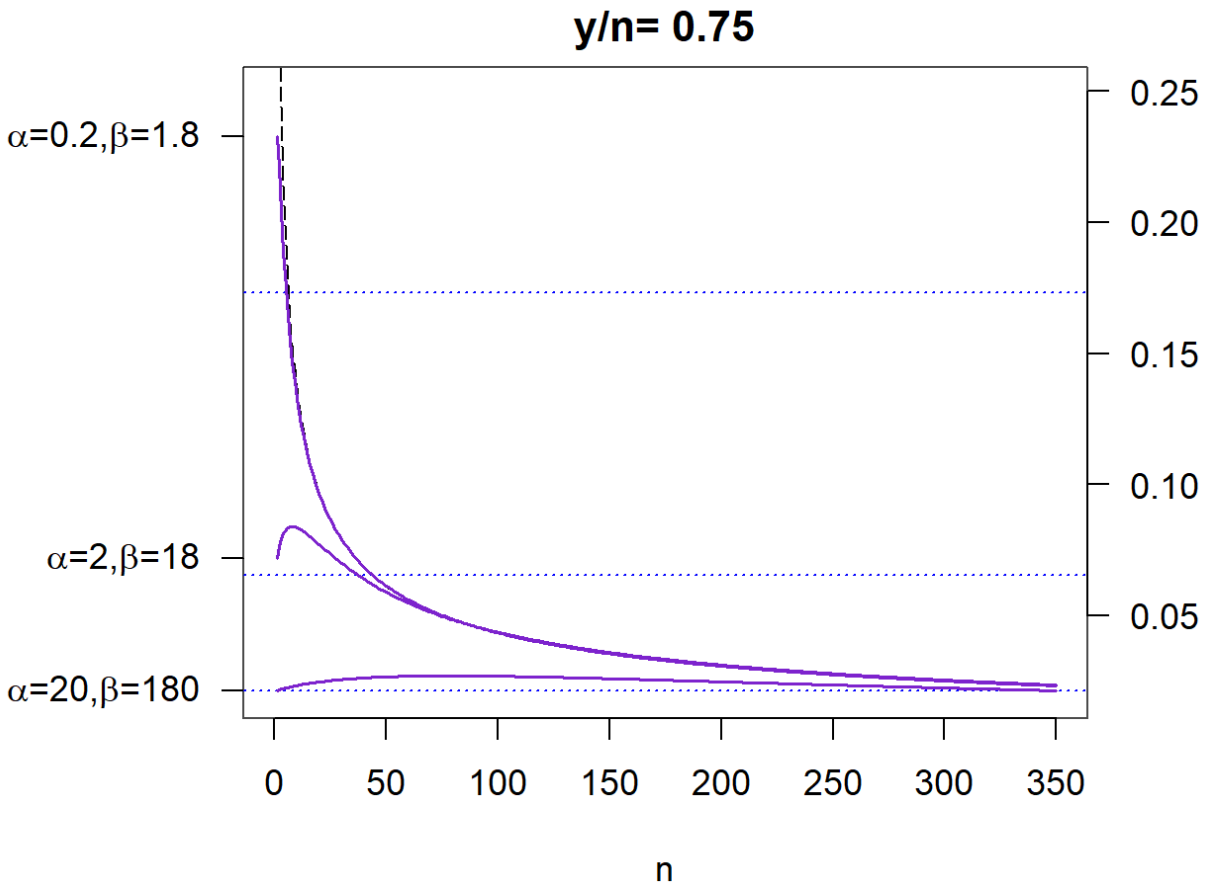



```
## sd(MLE) with n=1:3,10,50,100,350  0.43 0.31 0.25 0.14 0.06 0.04 0.02
## prior sd: 0.289 0.109 0.035

## post sd with n=1:3,10,50,100,350  0.25 0.22 0.19 0.13 0.06 0.04 0.02
## post sd with n=1:3,10,50,100,350  0.11 0.1 0.1 0.09 0.06 0.04 0.02
## post sd with n=1:3,10,50,100,350  0.04 0.04 0.04 0.03 0.03 0.03 0.02
```

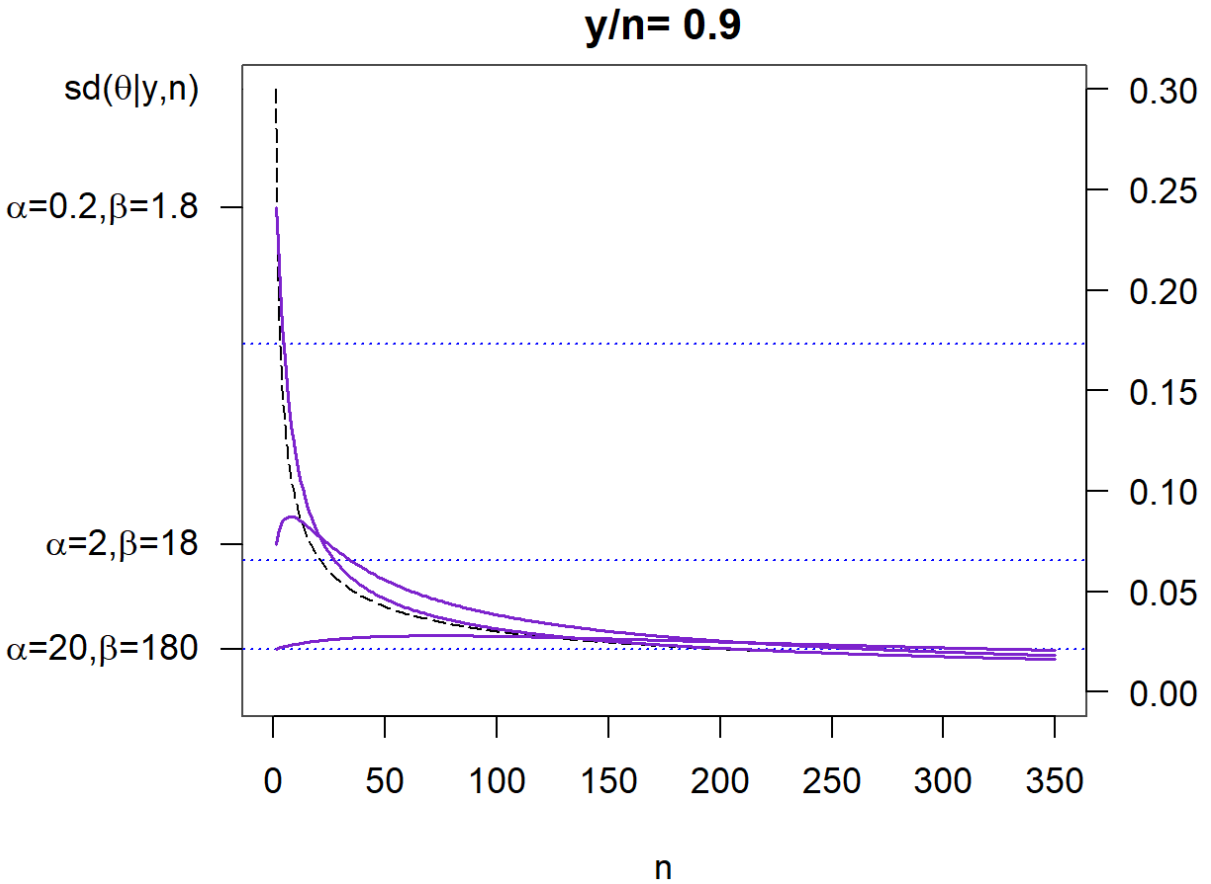
Posterior variance: conflicting priors

Posterior variance as n increases for different priors

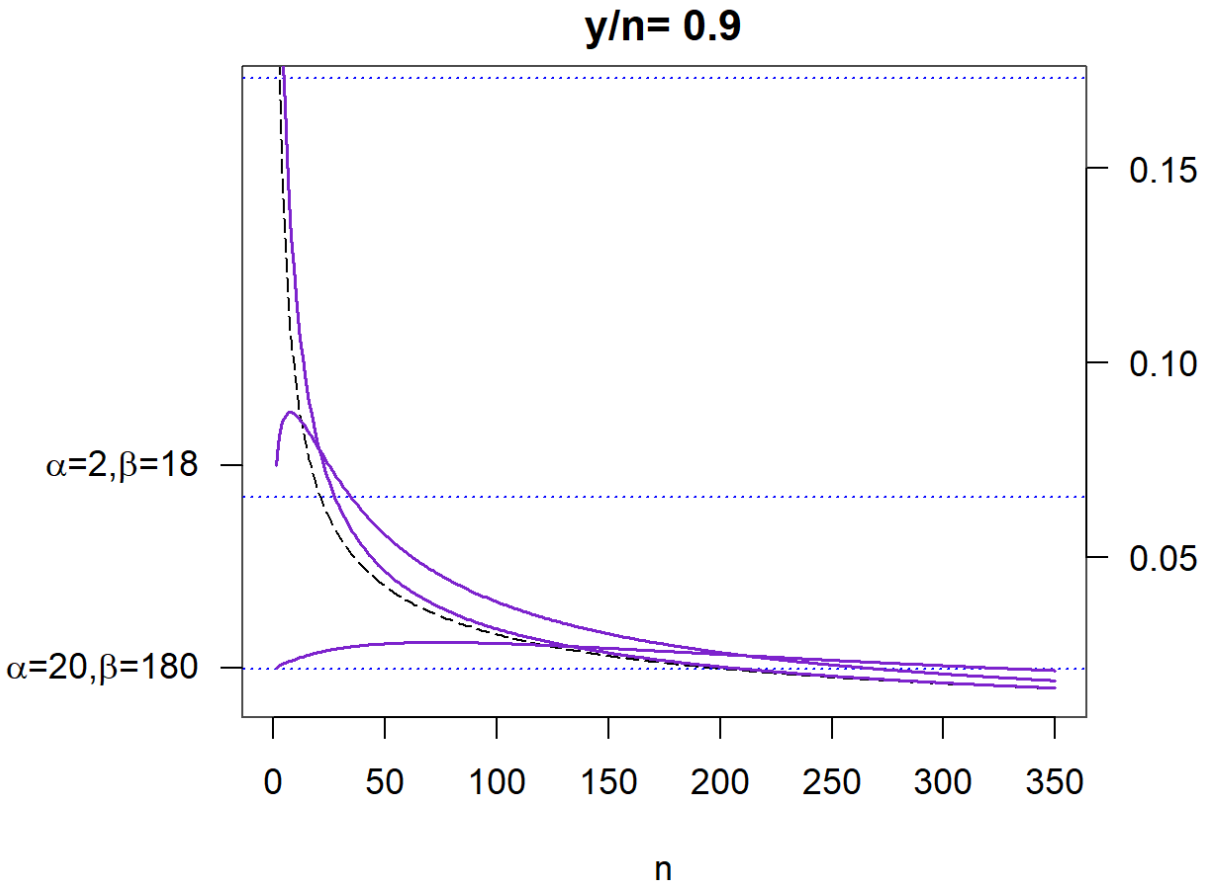





```
## sd(MLE), n=1:3,10,50,100,350  0.433 0.306 0.25 0.137 0.061 0.043 0.023
## prior sd: 0.173 0.065 0.021

## post sd, n=1:3,10,50,100,350  0.233 0.221 0.204 0.133 0.061 0.043 0.023
## post sd, n=1:3,10,50,100,350  0.072 0.076 0.079 0.084 0.059 0.044 0.023
## post sd, n=1:3,10,50,100,350  0.021 0.022 0.022 0.023 0.027 0.027 0.021
```




```
## sd(MLE), n=1:3,10,50,100,350  0.3 0.212 0.173 0.095 0.042 0.03 0.016

## prior sd: 0.17 0.07 0.02

## post sd, n=1:3,10,50,100,350  0.241 0.224 0.201 0.117 0.046 0.032 0.016
## post sd, n=1:3,10,50,100,350  0.074 0.079 0.082 0.087 0.056 0.038 0.018
## post sd, n=1:3,10,50,100,350  0.021 0.022 0.022 0.024 0.028 0.028 0.021
```




```
## sd(MLE), n=1:3,10,50,100,350  0.3 0.212 0.173 0.095 0.042 0.03 0.016

## prior sd: 0.17 0.07 0.02

## post sd, n=1:3,10,50,100,350  0.241 0.224 0.201 0.117 0.046 0.032 0.016
## post sd, n=1:3,10,50,100,350  0.074 0.079 0.082 0.087 0.056 0.038 0.018
## post sd, n=1:3,10,50,100,350  0.021 0.022 0.022 0.024 0.028 0.028 0.021
```

A priori non-informative, a priori proprie e improprie

- Vaghe, piatte, diffuse o ***noninformative***
  - tendono a "to let the data speak for themselves"
  - tipicamente si definiscono noninformative distribuzioni che sono *flat* sopra l'intero asse reale (e.g., $\propto 1$). A priori noninformative comuni includono distribuzioni uniformi *larghe* (e.g. $U(-1000, 1000)$ o $U(0, 1000)$ per parametri solo positivi) o distribuzioni normali diffuse (e.g. $N(0, 10000)$)
  - *flat* non è sempre vero che sia non-informativa
    - Assigning flat prior distributions to transformed parameters often yields highly skewed, strongly informative priors for the parameter in the original scale.
  - *flat* può essere una scelta stupida
    - A more accurate definition of noninformative priors would be ‘distributions that possess a range of uncertainty larger than any plausible parameter value’
  - rendendo la a priori *flat* in qualche parte si può renderla *non-flat* in qualche altra parte

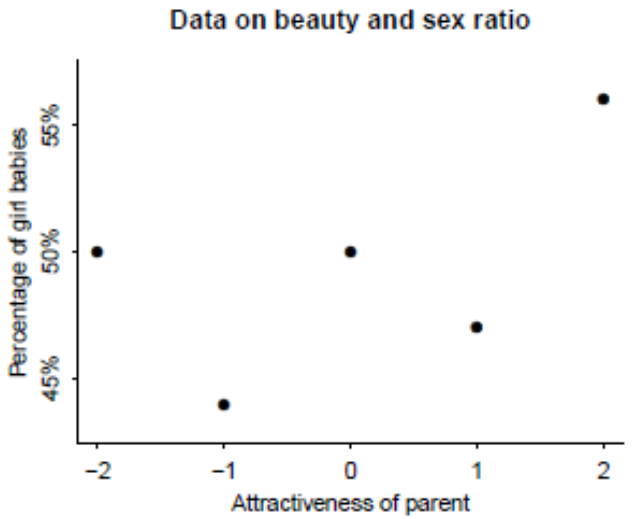
- a priori proprie hanno $\int p(\theta) = 1$
- la densità delle a priori improprie non ha un integrale finito
  - la a posteriori può ancora talvolta essere proria

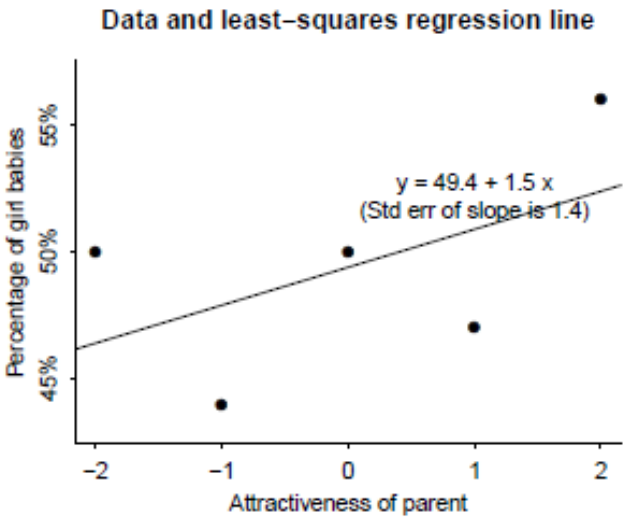
A priori debolmente informative

- Le a priori ***debolmente informative*** producono un comportamento delle a posteriori migliore dal pdv computazionale
    - abbastanza spesso c'è almeno una certa conoscenza della scala
    - utile anche se si hanno informazioni dalle osservazioni precedenti, ma non siamo certi di quanto queste informazioni siano applicabili all'incertezza del nuovo caso
- Costruzione
    - Inizia con una qualche versione di distribuzione a priori non informativa e quindi **aggiungi** informazioni sufficienti a rendere le inferenze "ragionevoli".
    - Inizia con una a priori forte e molto informativa e **amplia**la per tenere conto dell'incertezza nelle proprie convinzioni a priori e nell'applicabilità di qualsiasi distribuzione a priori basata sulla storia passata a nuovi dati.
- Raccomandazioni sulla scelta preliminare del team di Stan https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

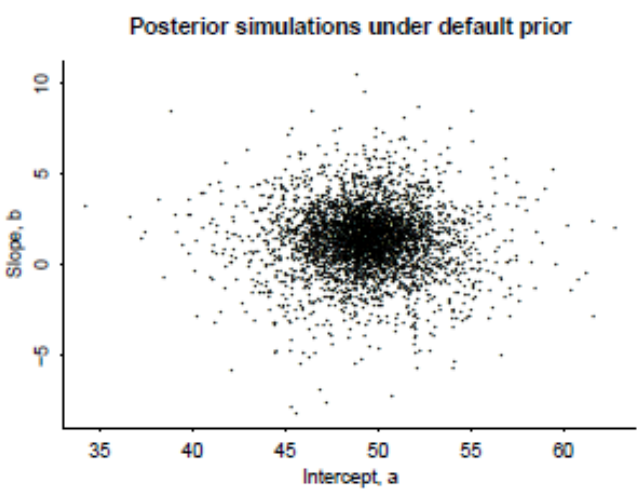
Esempio di a priori informativa

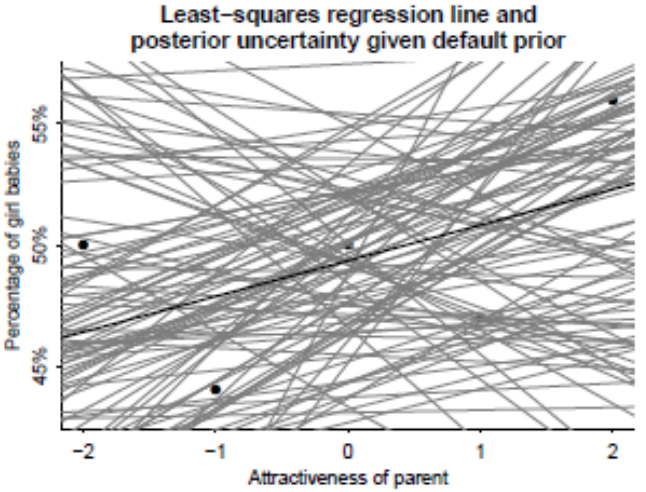
- La percentuale di nascite femminil è segnatamente stabile intorno a 48.5% , raramente varia di più di 0.5% da questo tasso
- C'è uno studio sulla percentuale di nascite di femmine tra genitori nelle categorie di attrattività 1–5 (valutate da intervistatori in un sondaggio faccia a faccia)

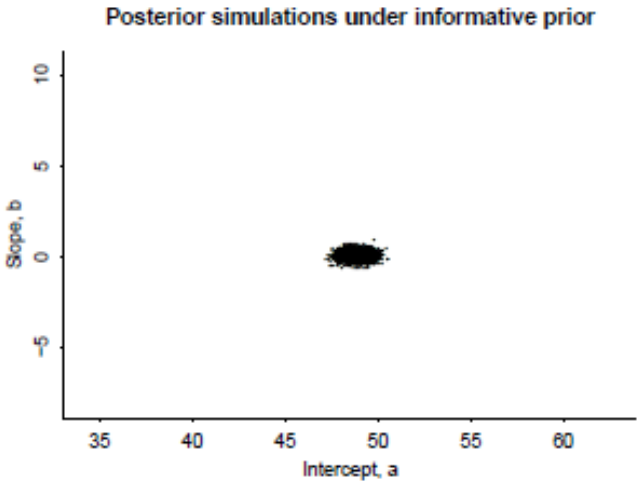


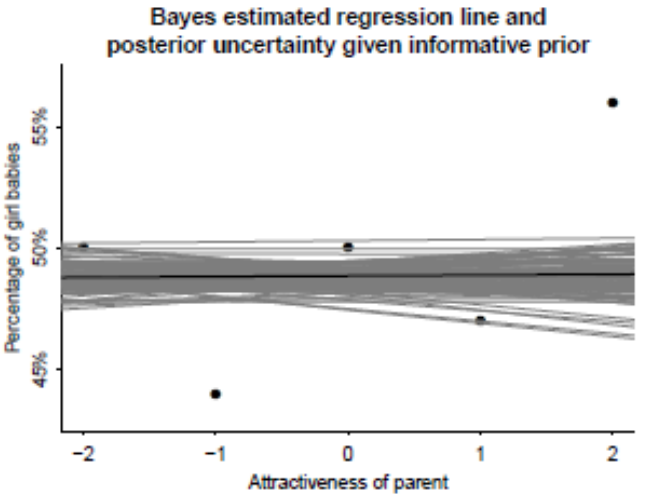


Esempio di a priori informativa









Stima della distribuzione a posteriori via simulazione

Soluzioni analitiche (fare calcoli matematici)

- Facilitate dalle a priori coniugate
  - Per $x \sim \text{Binom}(\theta, n), \theta \sim \text{Beta} \rightarrow \theta|x \sim \text{Beta}$
- Not feasible in more complex models

Quando non possiamo ottenere analiticamente l'a posteriori, si rende necessario stimarla o approssimarla in qualche modo

- Si può approssimarla o stimarla
- Si vuole un approccio flessibile e generale alla stima delle distribuzioni
  - Realizzato tramite simulazione
  - Un po' adesso, di più avanti

Stima basata sulla simulazione

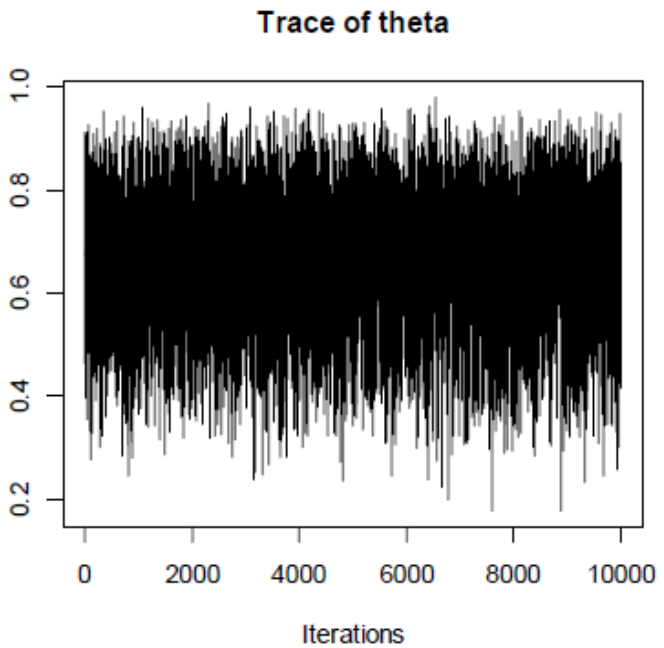
Si costruisce un algoritmo di campionamento per ***simulare*** o ***estrarre dalla*** a posteriori. Si campiona molte di queste estrazioni, che servono ad approssimare empiricamente la distribuzione a posteriori, e può essere utilizzato per approssimare empiricamente le statistiche riassuntive.

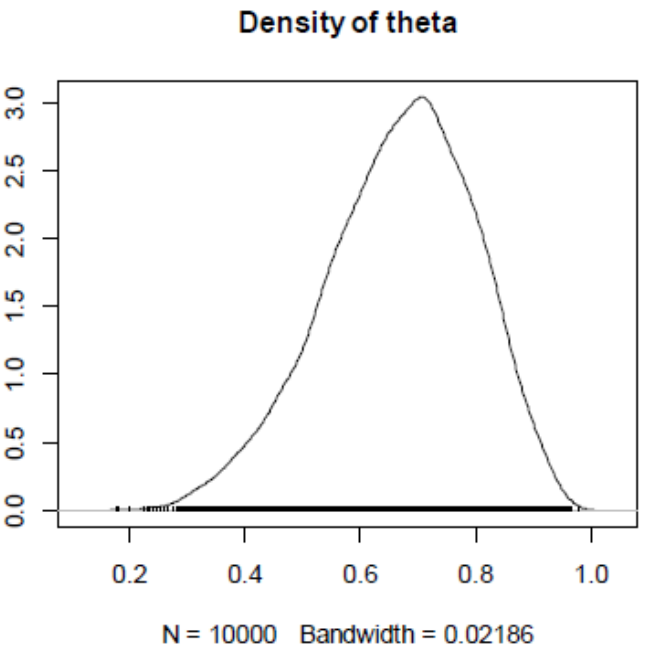
Principio di Monte Carlo:

> Anything we want to know about a random variable θ can be learned by sampling many times from $f(\theta)$, the density of θ.
>
> -- Jackman (2009, p. 133)

Simulazione dell'a posteriori nel modello Beta-Binomiale

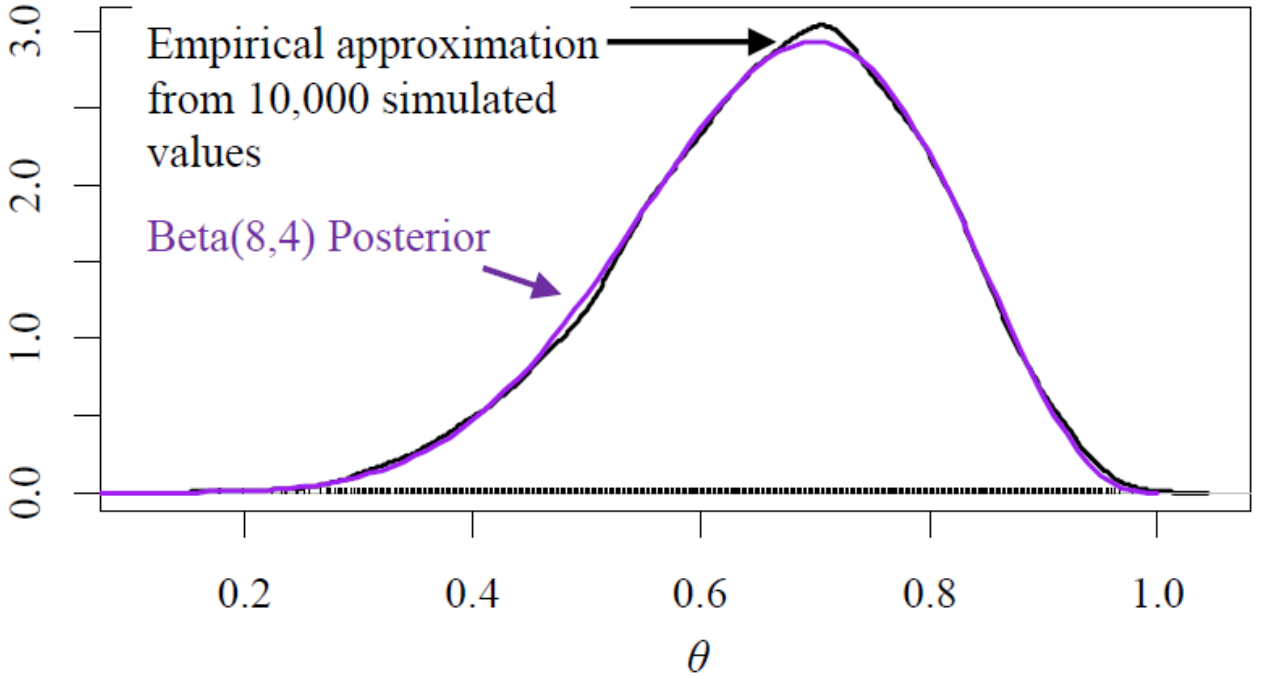
Si costruisce un algoritmo di campionamento per ***simulare*** o ***estrarre dalla*** a posteriori.





Modello Beta-Binomiale: densità

Si colleziona molte di queste estrazioni, che servono ad approssimare empiricamente la distribuzione a posteriori



Modello Beta-Binomiale: statistiche riassuntive

ed anche ad approssimare empiricamentele statistiche riassuntive.

```
Iterations = 1:10000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000
1. Empirical mean and standard deviation for each
   variable, plus standard error of the mean:
  Mean             SD          Naive  SE   Time-series SE
0.667688       0.130130       0.001301       0.001272

2. Quantiles for each variable:
  2.5%      25%         50%         75%         97.5%
0.3874      0.5815      0.6779      0.7636      0.8925
```