One-parameter models

General approach to Bayesian data modelling, Binomial model

Teacher: Matilde Trevisani

DEAMS

A.A. 2024/2025

(aggiornato: 2025-04-01)

Agenda (about 3 lectures)

One-parameter models

- General approach to Bayesian data modelling
- A first example
- Note on accumulation of evidence
- Binomial model
- Note on impact of more evidence
- Summarizing posterior distributions
- Conjugacy
- Interplay between priors and data
- Normal model
- Poisson model
- Other models

General approach to Bayesian data modelling

A ***Bayesianly justifiable*** analysis is one that

> “treats known values as observed values of random variables, treats unknown values as unobserved random variables, and calculates the conditional distribution of unknowns given knowns and model specifications using Bayes’ theorem.”
>
> -- *Rubin (1984, p. 1152)*

3-Step General Approach to Bayesian Modeling

1. Set up the **full probability model**: the joint distribution of all entities, including observables (y) and unobservables (θ) in accordance with all that is known about the problem

$$p(y, \theta) \propto p(y|\theta)p(\theta)$$

   - $p(y|\theta) \propto L(\theta) = (L(\theta, y))$ is the model for the conditional probability of the data, that is (proportional to) the **likelihood**
   - $p(\theta)$ is the **prior** distribution for the unknown parameters, reflecting what is believed about the situation

2. Condition on the observed data (x), calculate the conditional probability distribution for the unobservable entities (θ) of interest given the observed data: the **posterior** distribution

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

3. Examine fit of the model, tenability/sensitivity of assumptions, reasonable conclusions?, respecify, summarize results, etc.

See ambiguous notation

Notation

The main characters:

- We denote with Greek letters, typically, θ, the parameter(s), unobservable quantities. θ can be a scalar or a vector.
- The observed data are denoted by y, if data are gathered on n units:

  $$y = (y_1, \ldots, y_n)$$

  where y_i can be a scalar or a vector (if more than one variable is observed on each unit). y can then be a scalar, a vector, or a matrix.
- We will also use unknown but potentially observable quantities, that is, future observations, these will be denoted as $\tilde{y}$.
- If covariates are available, these will be denoted by x.

Model Specification

Specifying a **Bayesian model** means specifying:

- The distribution of y conditional on the parameter $\theta: y|\theta \sim p(y|\theta)$
- The prior distribution on $\theta: \theta \sim p(\theta)$

Putting these together, we have specified the **joint distribution of** (y, θ):

$$p(y, \theta) = p(y|\theta)p(\theta)$$

and we can obtain the **marginal distribution of** y as:

$$p(y) = \int_\Theta p(y, \theta) d\theta = \int_\Theta p(y|\theta)p(\theta) d\theta$$

Posterior distribution

Inference on θ will be based on the posterior distribution, which is derived through a straightforward application of Bayes theorem

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

The posterior distribution contains all the information on θ we have (from the data and prior to observing the data).

The work will have to do is to understand

- how to summarize the information in $p(\theta|y)$, to obtain for instance point and interval estimates or to perform hypotheses testing;
- how to explore the distribution, but for simple examples $p(y)$ is difficult to derive (impossible to derive analytically), so exploration of the posterior will be based on computational machinery (MCMC and other stuff) whose starting point is

$$\pi(\theta|y) \propto p(y|\theta)p(\theta)$$

Predictive Distribution

We are sometimes interested in "unknown but potentially observable quantities" $\tilde{y}$ (e.g., prediction of y on new statistical units).

We assume that they behave like the data y, that is:

$$\tilde{y}|\theta \sim p(\tilde{y}|\theta)$$

Hence, unconditionally, the distribution of $\tilde{y}$ is:

$$p(\tilde{y}) = \int_{\Theta} p(\tilde{y}|\theta)p(\theta)d\theta$$

which is the same as y. This is also called the ***prior predictive distribution***. After the data y have been observed, we can compute the ***posterior predictive distribution***:

$$p(\tilde{y}|y) = \int_{\Theta} p(\tilde{y},\theta|y)d\theta = \int_{\Theta} p(\tilde{y}|\theta,y)p(\theta|y)d\theta = \int_{\Theta} p(\tilde{y}|\theta)p(\theta|y)d\theta$$

where we note that the conditional iid assumption implies that:

$$p(\tilde{y}|\theta,y) = p(\tilde{y}|\theta).$$

Exchangeability

A common hypothesis in statistical inference is that observations are independent and identically distributed (iid), meaning we collect $y_1, \ldots, y_n$ and assume these are iid.

In Bayesian inference, where the inference process is fully probabilistic. independence of observations would imply that we cannot learn about future observations from past ones (since y_{n+1} would be independent of $y_1, \ldots, y_n$).

Instead, we assume observations are ***exchangeable***, meaning the joint distribution of $(y_1, \ldots, y_n)$ is invariant to index permutations:

$$p(y_1, \ldots, y_n) = p(y_{i_1}, \ldots, y_{i_n})$$

for any permutation $(i_1, \ldots, i_n)$ of $(1, \ldots, n)$.

Exchangeability and Conditional Independence

We will usually specify the model assuming that

- $y_1, \ldots, y_n$ are iid conditional on θ
- $\theta \sim p(\theta)$

This **implies** that $y_1, \ldots, y_n$ are exchangeable. In fact, consider the unconditional distribution:

$$
\begin{aligned}
p(y_{i_1}, \ldots, y_{i_n}) &= \int p(y_{i_1}, \ldots, y_{i_n} | \theta) p(\theta)\, d\theta \\
&= \int \prod_{j=1}^{n} p(y_{i_j} | \theta) p(\theta)\, d\theta \\
&= \int \prod_{i=1}^{n} p(y_i | \theta) p(\theta)\, d\theta = p(y_1, \ldots, y_n)
\end{aligned}
$$

de Finetti's Theorem

For **binary** variables $y_1, \ldots, y_n$, exchangeability is equivalent to conditional iid:

Theorem (de Finetti): Let $Y_1, Y_2, \ldots, Y_n, n \to \infty$, be a sequence of Bernoulli r.v., then they are exchangeable if and only if there exists a random variable θ valued in $[0, 1]$ such that:

$$p(y_1, \ldots, y_n) = \int_0^1 \theta^{\sum y_i} (1-\theta)^{n - \sum y_i} dP(\theta).$$

An extension of this theorem exists for **general** random variables.

Non independence as θ unknown

The following are equivalent:

- $y_1, \ldots, y_n$ are exchangeable.
- $y_1, \ldots, y_n$ are iid conditional on θ.

This means:

> Observations are IID if we know the data-generating mechanism.

Since we do not know it, observations are not independent. Instead:

> y_1 gives information about y_2 ***because*** it provides information about the data-generating mechanism θ.

More on Bayesian prediction interpretation

("The usual Bayesian story":) Bayesian statistics is often described as consisting of assigning a prior on θ and using Bayes rule to compute the posterior distribution. Obtaining the predictive distribution,

$$p(\tilde{y}|y) = \int_{\Theta} p(\tilde{y}|\theta, y) dp(\theta|y)$$

is then just a matter of computations. Bayesian statistics is deeper than that! And a first basic concept we should recall is the interpretation of the **Bayesian predictive distribution**.

Bayesian statistics is about acting under uncertainty, or incomplete *information* (from the data, from domain knowledge, etc.).
If probability is the prescribed formal language to describe this (incomplete) information, then the evolution of information, or *learning*, is expressed through *conditional probabilities*.
In particular, learning on the next observation based on the observed is expressed through the conditional distribution $p(\tilde{y}|y)$.
This leads us to the interpretation of the Bayesian predictive distribution:

it is a **learning rule** that formalizes, through conditional probability, how we learn about future events given the available information.

(Thus, it is not meant as the ‘physical mechanism’ generating $\tilde{Y}$ given the past, like in the classic setting).

Exchangeability with known model parameters

For the following scenarios, answer:

1. Are y_1 and y_2 exchangeable?
2. Are they independent?
3. Can we act as if they are independent?

Case A: A box has one black and one white ball. Pick y_1 randomly, replace it, then pick y_2.

Case B: A box has one black and one white ball. Pick y_1 randomly, do *not* replace it, then pick y_2.

Case C: A box has a million black and a million white balls. Pick y_1 randomly, do *not* replace it, then pick y_2.

A similar set of questions follows when the exact number of black and white balls is unknown.

Exchangeability with unknown model parameters

For the following scenarios, answer:

1. Are y_1 and y_2 exchangeable?
2. Are they independent?
3. Can we act as if they are independent?

Case A: A box has n black and white balls, but we don't know how many of each color. Pick y_1 randomly, replace it, then pick y_2.

Case B: A box has n black and white balls, but we don't know how many of each color. Pick y_1 randomly, do *not* replace it, then pick y_2.

Case C: Same as B but we know that there are many balls of each color in the box.

Exchangeability and Independence

You don’t need to understand the term exchangeability before learning Hierarchical Bayesian Models (Chapter 5).

At this point,

- we consider exchangeable models ***for data***, $y_1, \ldots, y_n$, in the form of likelihoods in which the n observations are iid, given some parameter vector θ. (Later we will consider exchangeability for parameters.)

- Exchangeability is less strict condition than independence.

    - independence implies exchangeability
    - exchangeability does not imply independence

- exchangeability is related to what information is available (instead of the properties of unknown underlying data generating mechanism. See slide on Bayesian prediction interpretation)

    - Often we may assume that observations are in fact dependent, but if we can’t get information about these dependencies we may assume those observations as exchangeable. "Ignorance implies exchangeability."

A first example

Inference about a discrete quantity

In what follows we consider a real example of the very simplest case of Bayesian calculation.

It is not typical of *statistical* applications of Bayesian inference, as it deals with the **estimation of a single individual's state** (gene carrier or not) - and a very small data sample, rather than with the estimation of a parameter that describes an entire population.

Both the estimand and the observed variable are binary.

Inference about a Genetic Status: Prior

Human males have one X-chromosome and one Y-chromosome, whereas females have two X-chromosomes, each chromosome being inherited from one parent.

Hemophilia is due to a recessive gene in the X-chromosome, that is, if X^* denotes an X-chromosome with the hemophilia gene,

- $X^* X^*$ is a female with the disease
- $X^* X$ is a female without the disease but with the gene
- $X^* Y$ is a male with the disease

Mary has

- an affected brother $\Rightarrow$ $X^* Y$
- an unaffected mother $\Rightarrow$ XX^* or XX
- an unaffected father $\Rightarrow$ XY

Overall, the mother must be XX^*.

Let $\theta = 1$ if Mary is a gene carrier (is XX^*) and 0 otherwise (XX), then *based on the above information*, **prior** to any observation,

$$P(\theta = 1) = \frac{1}{2}$$

Inference about a Genetic Status: Data Model and Likelihood

Data consist of the status of Mary's two sons, who are not affected.

Let then y_i be an indicator equal to 1 if the i-th son is affected:

$$P(y_i = 1|\theta) = \begin{cases} 0.5 & \text{if } \theta = 1 \\ 0 & \text{otherwise} \end{cases}$$

The outcomes of the two sons are exchangeable and, conditional on the unknown θ, are independent; we assume the sons are not identical twins.

The likelihood function corresponding to Mary's two sons is:

$$L(\theta) = P(y_1 = y_2 = 0|\theta) = \begin{cases} 0.25 & \text{if } \theta = 1 \\ 1 & \text{if } \theta = 0 \end{cases}$$

Inference about a Genetic Status: Prior Predictive

Data consist of the status of Mary's two sons, who are not affected.

We know that

$$P(y_1 = y_2 = 0|\theta) = \begin{cases} 0.25 & \text{if } \theta = 1 \\ 1 & \text{if } \theta = 0 \end{cases}$$

Let $y = (y_1 = y_2 = 0)$, the predictive probability is

$$\begin{aligned} P(y) &= P(y|\theta = 1)P(\theta = 1) + P(y|\theta = 0)P(\theta = 0) \\ &= 0.25 \times 0.5 + 1 \times 0.5 = 0.625 \end{aligned}$$

Inference about a Genetic Status: Posterior

Prior and likelihood are combined to obtain the posterior, let $y = (y_1 = y_2 = 0)$,

$$\begin{aligned} P(\theta = 1|y) &= \frac{P(y|\theta = 1)P(\theta = 1)}{P(y)} \\ &= \frac{P(y|\theta = 1)P(\theta = 1)}{P(y|\theta = 1)P(\theta = 1) + P(y|\theta = 0)P(\theta = 0)} \\ &= \frac{0.25 \times 0.5}{0.25 \times 0.5 + 1 \times 0.5} = 0.20 \end{aligned}$$

Intuitively it is clear that if a woman has unaffected children, it is less probable that she is a carrier.

When the parameter is discrete, the results can also be effectively described in terms of prior and posterior odds.
The posterior odds are given by the likelihood ratio times the prior odds: $\frac{p(\theta_1|y)}{p(\theta_2|y)} = \frac{p(y|\theta_1)}{p(y|\theta_2)} \frac{p(\theta_1)}{p(\theta_2)}$

$$\frac{0.2}{0.8} = \frac{P(\theta = 1|y)}{P(\theta = 0|y)} = \frac{P(y|\theta = 1)}{P(y|\theta = 0)} \frac{P(\theta = 1)}{P(\theta = 0)} = \frac{0.25}{1} \times 1$$

Inference about a Genetic Status: Predictive distributions

Prior to the observations the predictive distribution is

$$\begin{aligned}P(y_1 = 1) &= P(y_1 = 1|\theta = 1)P(\theta = 1) + P(y_1 = 1|\theta = 0)P(\theta = 0)\\ &= 0.5 \times 0.5 + 0 \times 0.5 = 0.25\end{aligned}$$

Given the data the posterior predictive is

$$\begin{aligned}P(\tilde{y}_3 = 1|y) &= P(\tilde{y}_3 = 1|\theta = 1, y)P(\theta = 1|y) + P(\tilde{y}_3 = 1|\theta = 0, y)P(\theta = 0|y)\\ &= P(\tilde{y}_3 = 1|\theta = 1)P(\theta = 1|y) + P(\tilde{y}_3 = 1|\theta = 1)P(\theta = 0|y)\\ &= 0.5 \times 0.2 + 0 \times 0.8 = 0.1\end{aligned}$$

Inference about a Genetic Status: Adding More Data

Suppose a third son is born and he is not affected, that is we have a new observation $y_3 = 0$, in order to obtain the new posterior distribution we can use the old posterior $P(\theta = 1|y)$ as a prior and update it based on the likelihood $P(y_3 = 0|\theta)$

$$\begin{aligned} P(\theta = 1|y, y_3 = 0) &= \frac{P(y_3 = 0|\theta = 1)P(\theta = 1|y)}{P(y_3 = 0|\theta = 1)P(\theta = 1|y) + P(y_3 = 0|\theta = 0)P(\theta = 0|y)} \\ &= \frac{0.5 \times 0.2}{0.5 \times 0.2 + 1 \times 0.8} = 0.111 \end{aligned}$$

A similar mechanism works with the odds

$$\begin{array}{ccccc} \dfrac{P(\theta = 1|y, y_3 = 0)}{P(\theta = 0|y, y_3 = 0)} & = & \dfrac{P(y_3 = 0|\theta = 1)}{P(y_3 = 0|\theta = 0)} & \dfrac{P(\theta = 1|y)}{P(\theta = 0|y)} \\ \dfrac{1}{8} & = & \dfrac{0.5}{1} & \dfrac{1}{4} \end{array}$$

The same result is obtained by starting from the prior and considering the data $y' = (y_1 = y_2 = y_3 = 0)$.

Sequential analysis

A key aspect of Bayesian analysis is the ease with which **sequential analyses** can be performed.

As new data arrives, we need updating the information.
Considering the whole data (y_1, y_2)

$$p(\theta|y_1, y_2) \propto p(y_1, y_2|\theta)p(\theta)$$

- Posterior distribution for θ given data y_1 and y_2
- Conditional distribution of y_1 and y_2 given θ
- Prior for θ

Assuming conditional independence, the likelihood can be partitioned:

$$p(y_1, y_2|\theta) = p(y_2|\theta)p(y_1|\theta)$$

$$\text{Then} \quad \begin{aligned} p(\theta|y_1, y_2) \propto p(y_1, y_2|\theta)p(\theta) &= p(y_2|\theta)p(y_1|\theta)p(\theta) \\ &\propto p(y_2|\theta)p(\theta|y_1) \end{aligned}$$

That is, $p(\theta|y_1, y_2)$ is partitioned into conditional distribution of the sole y_2 given θ and posterior distribution for θ given y_1 (up to a constant of proportionality)

Bayes' Theorem: Accumulation of Evidence

Dataset 1: $p(\theta|y_1) \propto p(y_1|\theta)p(\theta)$

Dataset 2: $p(\theta|y_1, y_2) \propto p(y_2|\theta)p(\theta|y_1)$

Dataset 3: $p(\theta|y_1, y_2, y_3) \propto p(y_3|\theta)p(\theta|y_1, y_2)$

Today's posterior is tomorrow's prior

Bayes' theorem as a mechanism for accumulating evidence

- Update diagnosis as symptoms, test results arrive
- Update beliefs about proficiency as students complete tasks
- Update beliefs about guilty as testimony is heard
- Do a study, use results as basis for prior for next study
- Makes Bayesian approach a natural framework for *meta-analysis* and related approaches that synthesize information from datasets

Binomial model

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.

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).

If we consider the trials *exchangeable* - we disregard the order - the data can be summarized by the total number of 1 (successes), which we denote by y.

Exchangeability is equivalent to say that conditional on θ, the probability of success in each trial, 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: $P(y_i = 1|\theta) = \theta \;\; \forall i$.

The sampling model for $y|\theta$ is then a binomial model

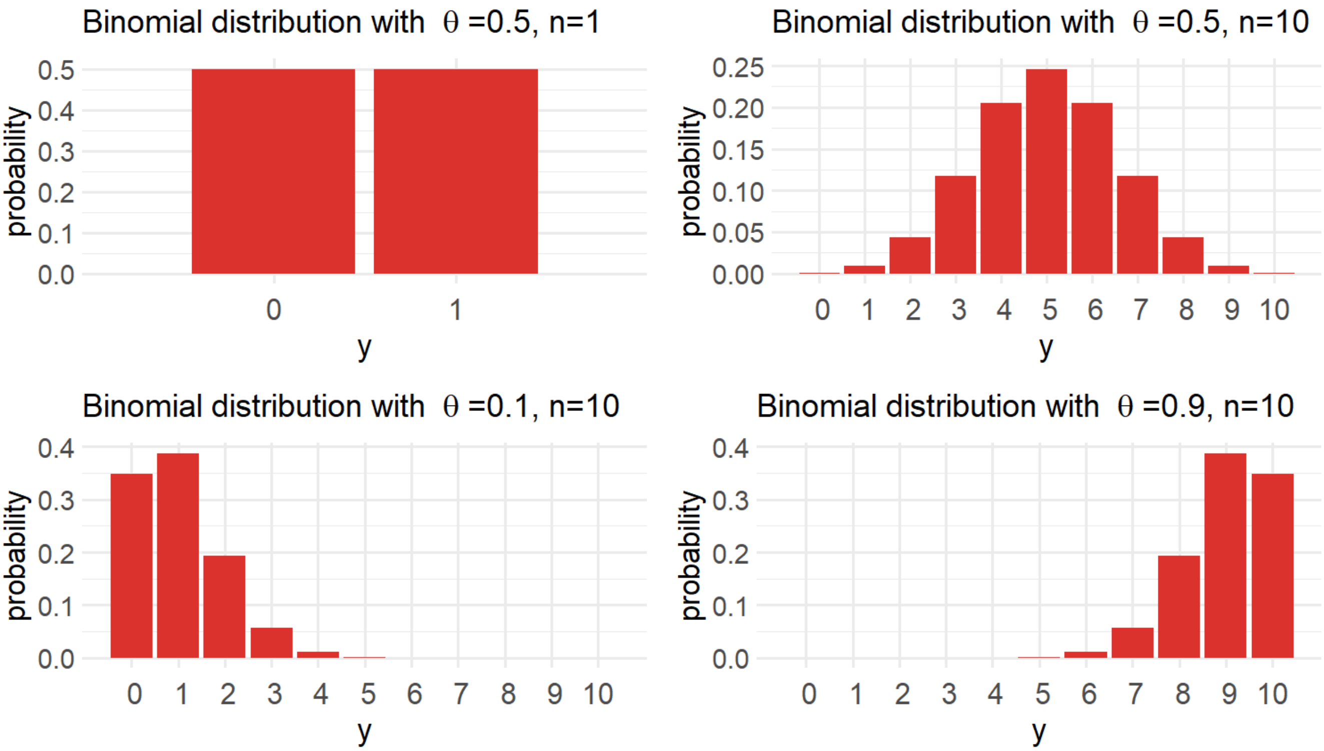
$$p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

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)

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

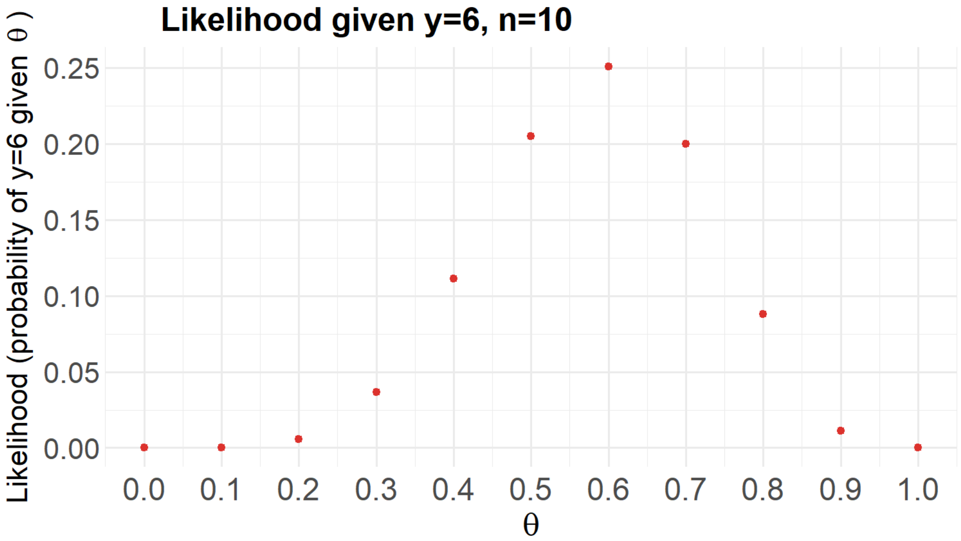


Binomial model as likelihood: θ unknown

Likelihood (continuous function of θ)

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

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



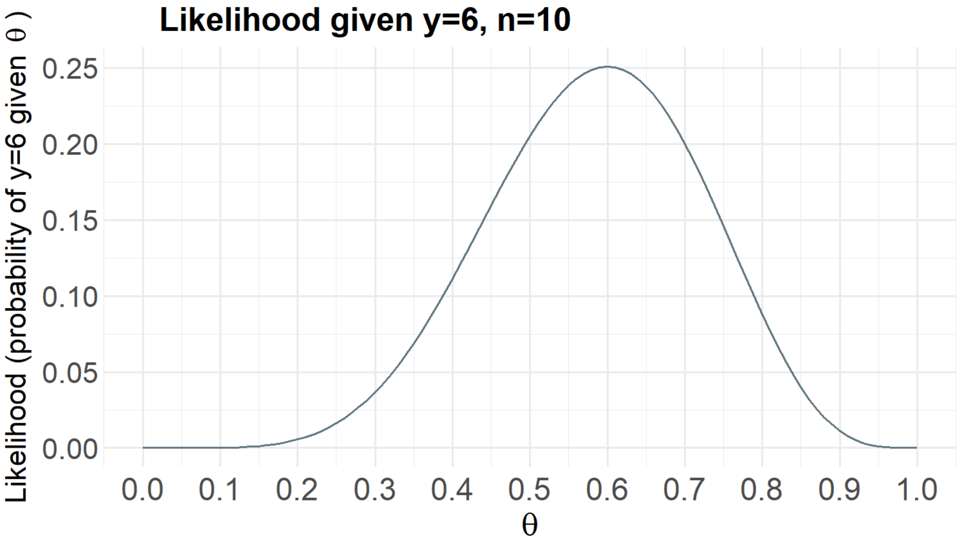
$p(y = 6|n = 10, \theta)$: 0.00 0.00 0.01 0.04 0.11 0.21 0.25 0.20 0.09 0.01 0.00

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

Let's start with a *uniform* prior

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

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

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

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

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 (constant given y) is the normalizing term

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

- Z has the form of the Beta function
  - when integrated on $(0,1)$ the result can be given with Gamma functions
  - with integers $\Gamma(n) = (n-1)!$
  - if integers are large this computation can be challenging and $\log\Gamma(\cdot)$ usually is computed in place of $\Gamma(\cdot)$

- If we compute it with $y = 6, n = 10$

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

  `gamma(y+1)*gamma(n-y+1)/gamma(n+2)` $\approx$ 0.0004329

Binomial model with uniform prior

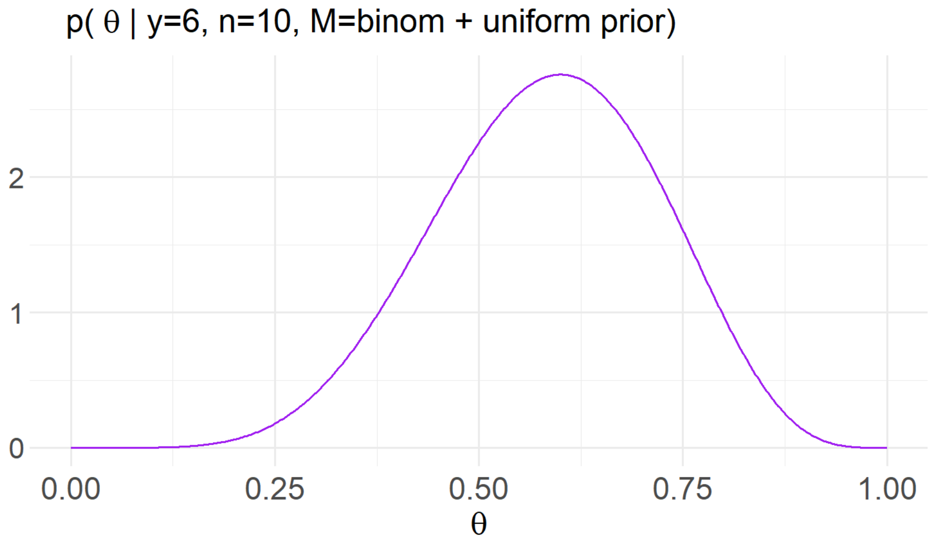
The **posterior**

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

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

The conditining to model M sometimes is shown explicitely

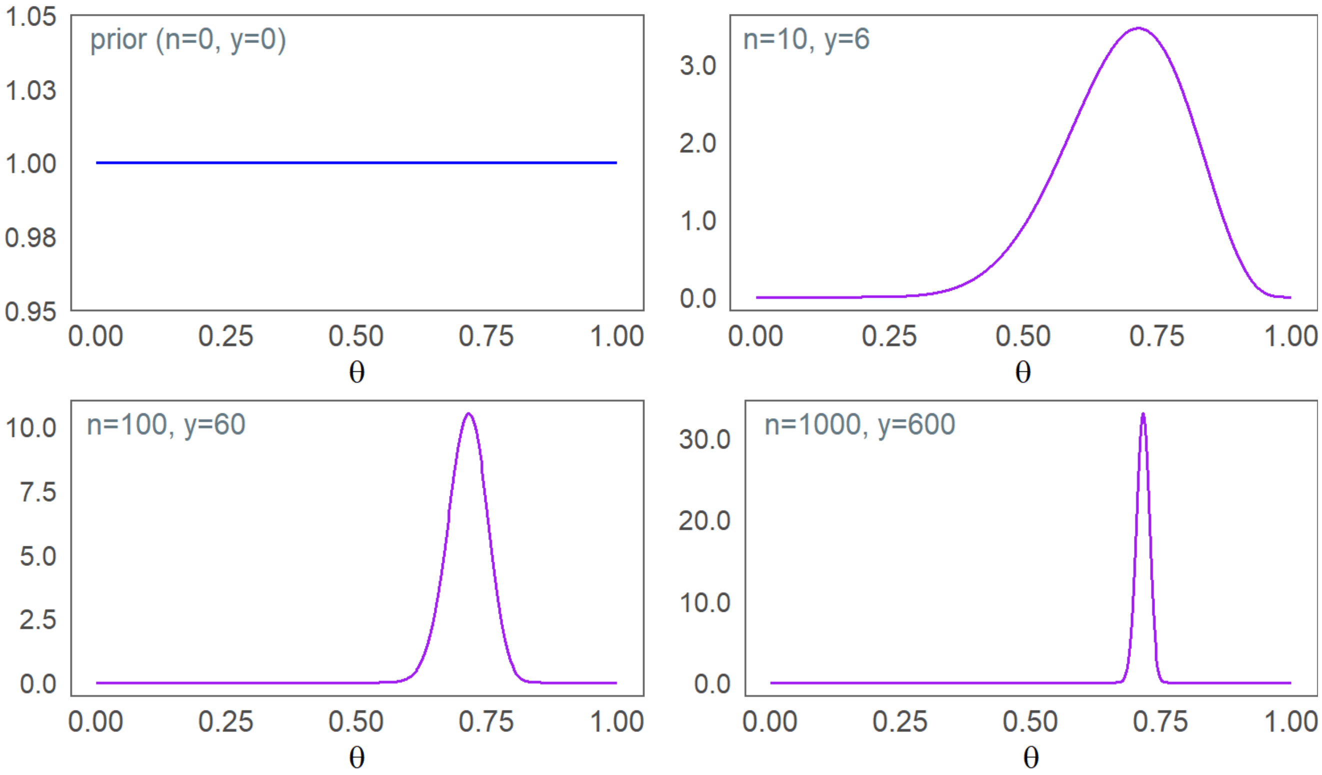
The posterior by Bayes rule is written

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

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

- makes clearer that the likelihood and the prior both constitute the model
- makes clearer that an absolute probability per $p(y)$ does not exist, but it depends on the model M
- in case of two models, we can evaluate the marginal probabilities $p(y|M_1)$ e $p(y|M_2)$
- It is usually implied to make the notation more concise.

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

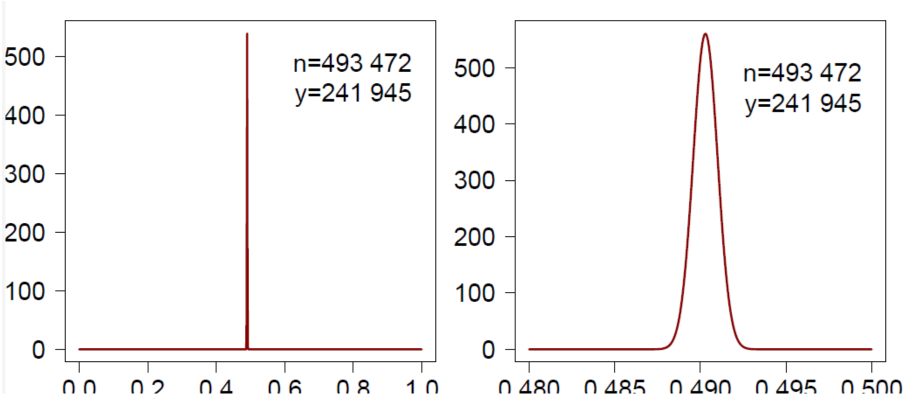
$$n = 241\,945 + 251\,527 = 493\,472; \quad y = 241\,945$$

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

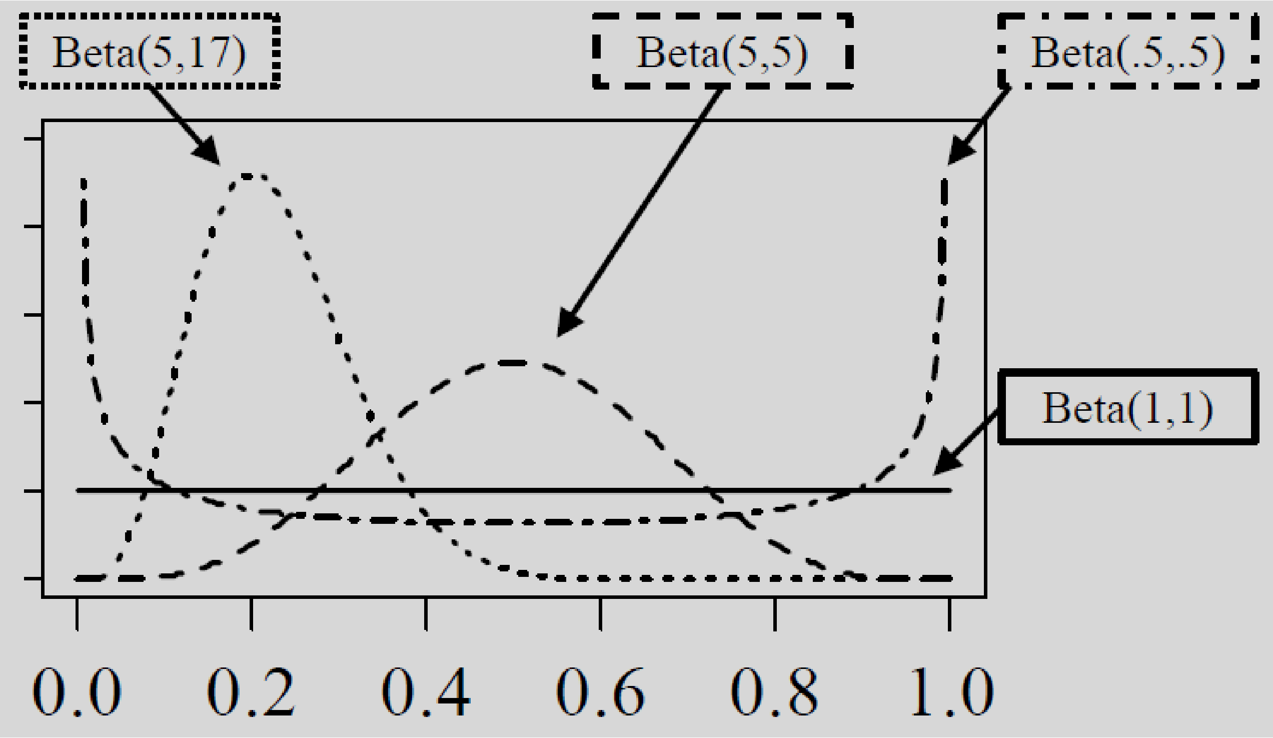


Binomial model: computation

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

- Beta CDF is not trivial to calculate
- E.g., `pbeta` in `R` uses a continued fraction with weighting factors and asymptotic expansion
- 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.

Beta distributions



Beta distributions

- Distribution on $[0, 1]$
- $\theta|\alpha, \beta \sim Beta(\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$
- The normalization constant is the reciprocal of

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

- Integral is finite if $\alpha, \beta > 0$, density is finite if $\alpha, \beta \geq 1$
  - Density has an asymptote in 0 if $\alpha < 1$, in 1 if $\beta < 1$.
- $E(Beta(\alpha, \beta)) = \frac{\alpha}{\alpha+\beta}$
- $\text{Var}(Beta(\alpha, \beta)) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
- $\text{Mode}(Beta(\alpha, \beta)) = \frac{\alpha-1}{\alpha+\beta-2}$ if $\alpha, \beta > 1, = 1$ if $\alpha \geq 1, \beta < 1, = 0$ if $\alpha < 1, \beta \geq 1$; has two modes if $\alpha, \beta < 1$
- $\alpha = E(\theta)(\alpha+\beta), \beta = (1 - E(\theta))(\alpha+\beta)$

Inference with binomial data

Classic/frequentist approach

Given θ, what are the probabilities of various possible risults for the r.v. y?

Weak Law of Large Numbers (Bernoulli theorem)

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

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

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

Bayesian approach

Given y, what are the probabilities of various possible risults for the r.v. 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

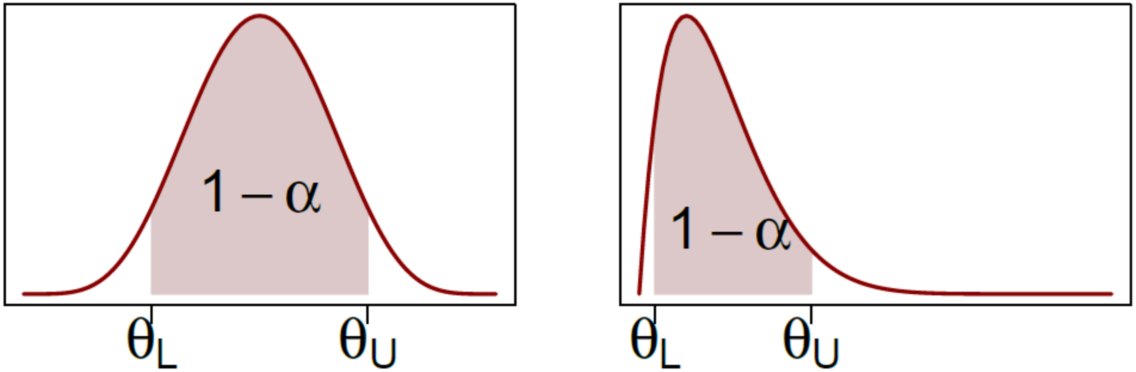
> Adopting an explicitly Bayesian approach would resolve a recurring source of confusion for these researchers, letting them say what they mean and mean what they say.
>
> -- Jackman (2009, p. xxviii)
>
> Frequentist CI theory says nothing at all about the probability that a particular, observed confidence 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

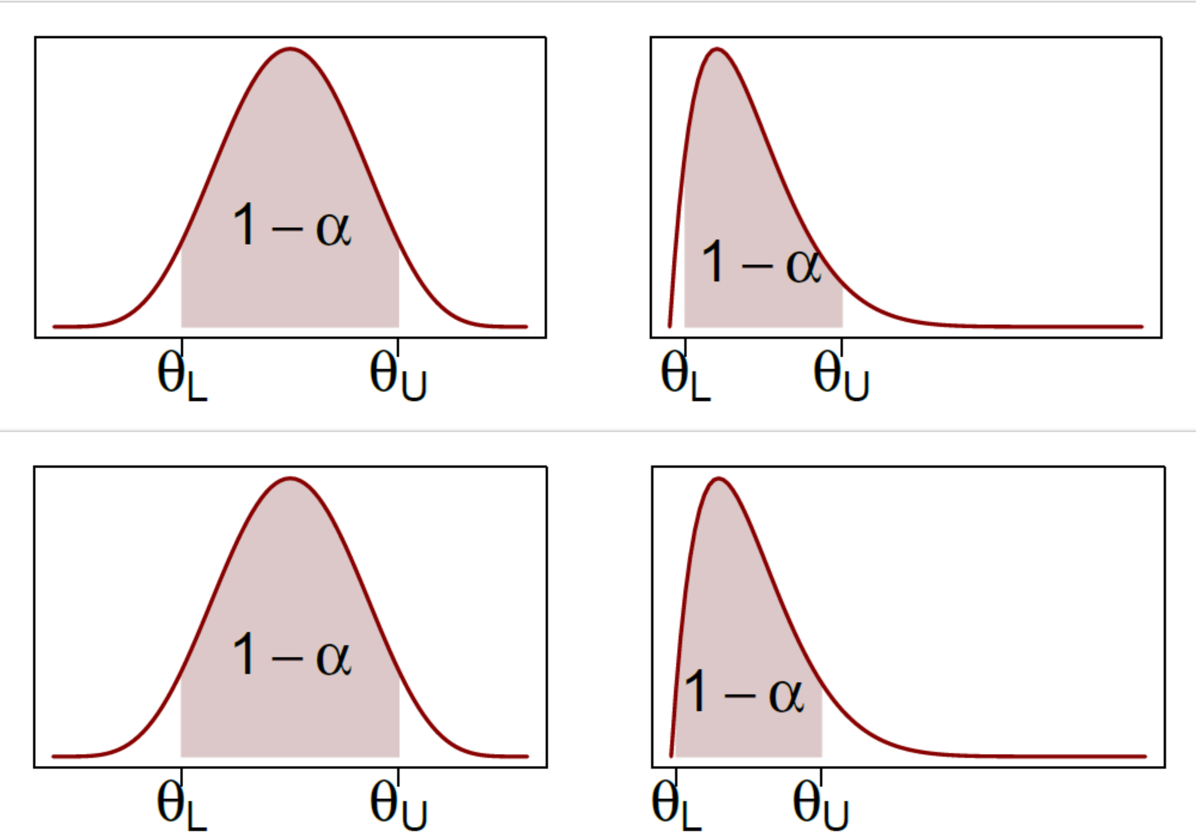
$$\{\theta|\pi(\theta|y) > c_\alpha\}$$

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

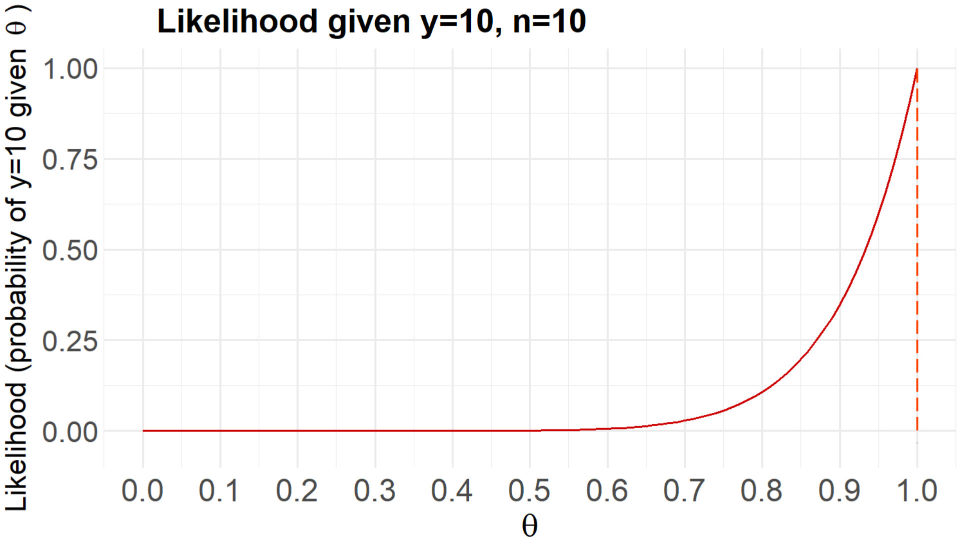
Benefits of integration

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

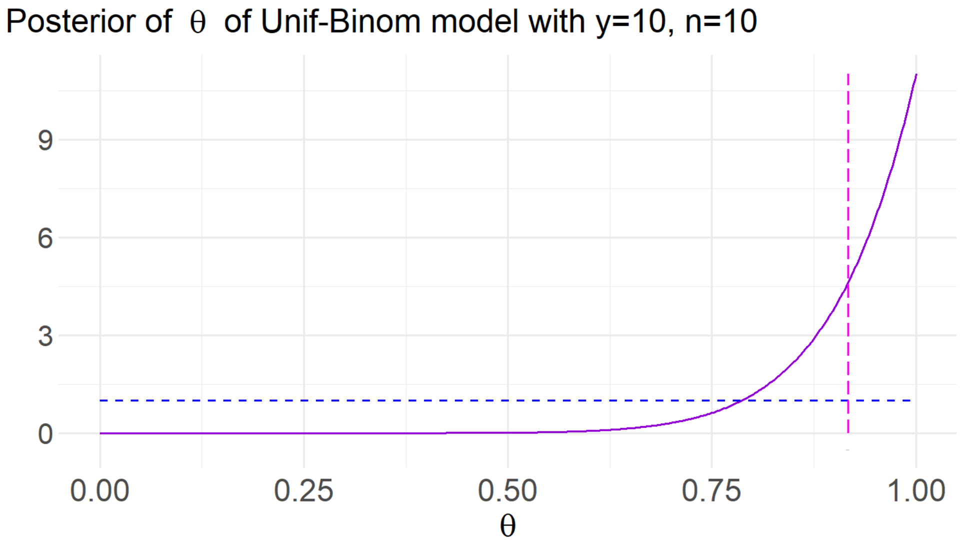
Perfect Response Pattern: ML



- 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?

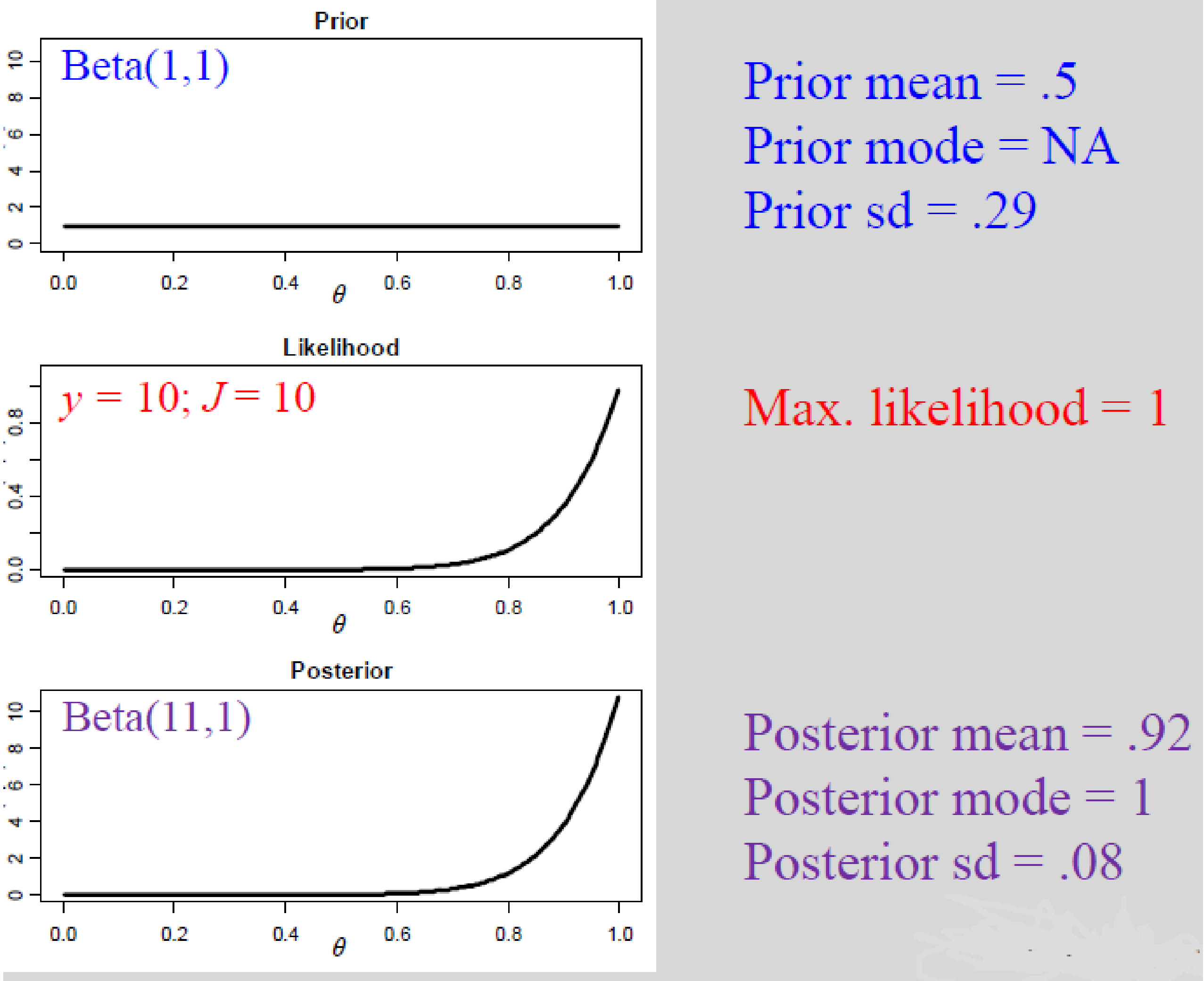
Perfect Response Pattern: Bayes with minimal prior information

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



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

Example: *Perfect Response Patterns*



Prior

Likelihood

Posterior

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 $y = (\sum_{i=1}^{n} y_i) \sim \mathrm{Bin}(n, \theta)$

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

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

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

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

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

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 $\beta = 1$

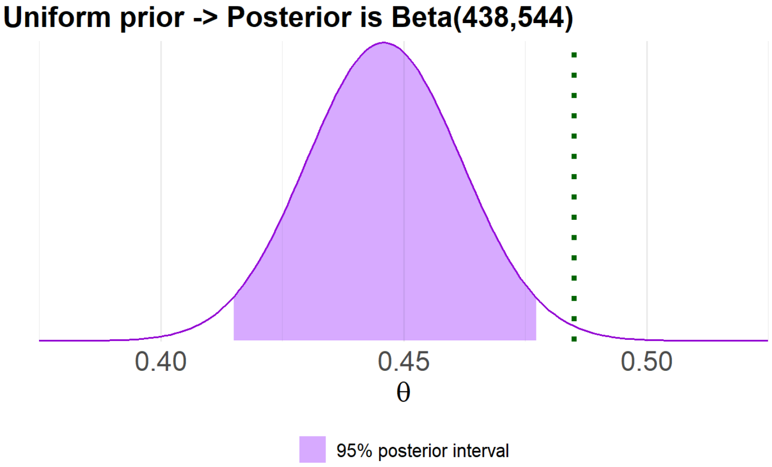
Example: placenta previa

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

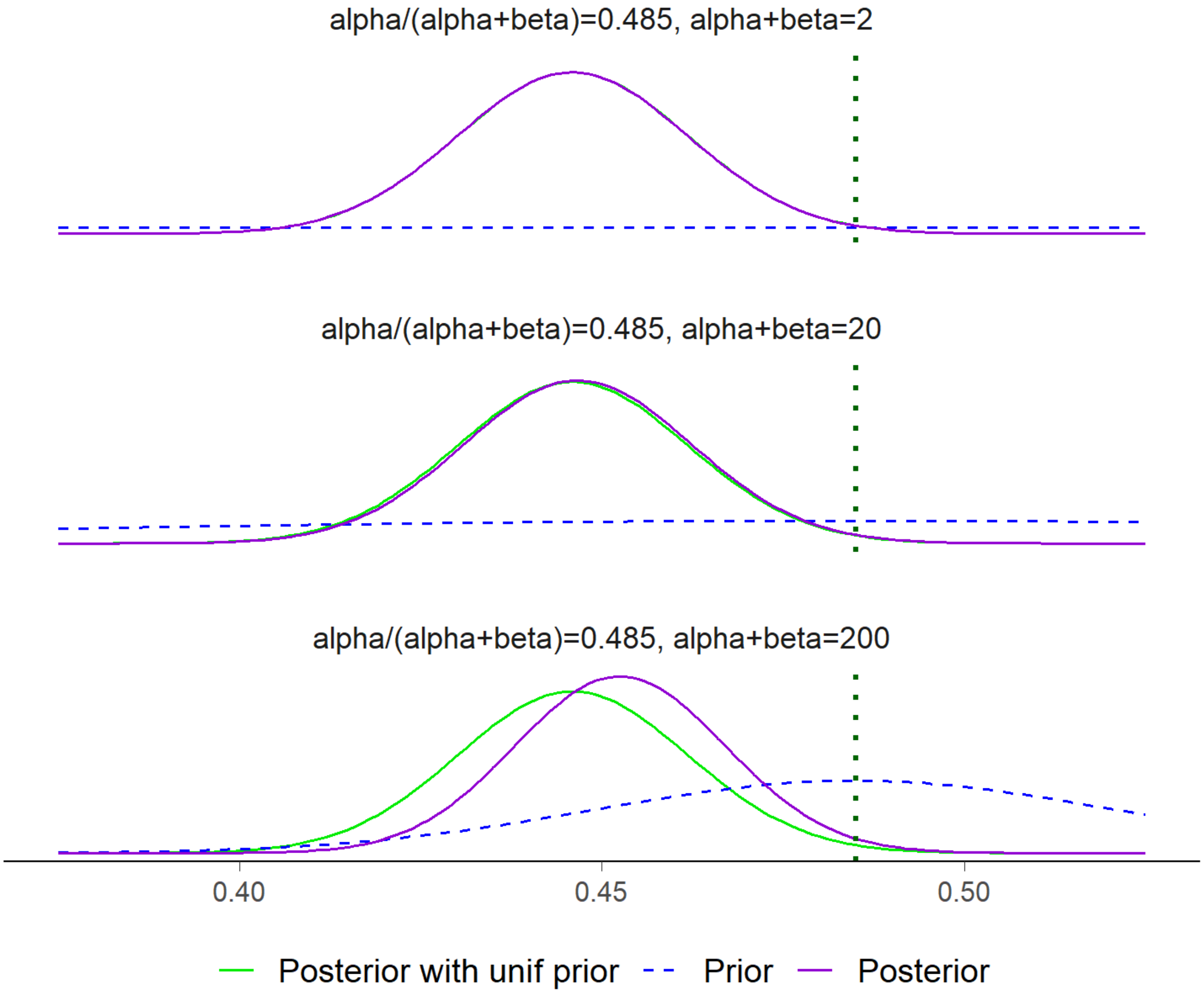
In a study in Germany, 437 females out of 980 births with placenta previa were observed.

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: $\propto \theta^{437}(1-\theta)^{980-437}$, $Beta(\theta|438,544)$
- $P(\theta < 0.485|y) = 0.9928$

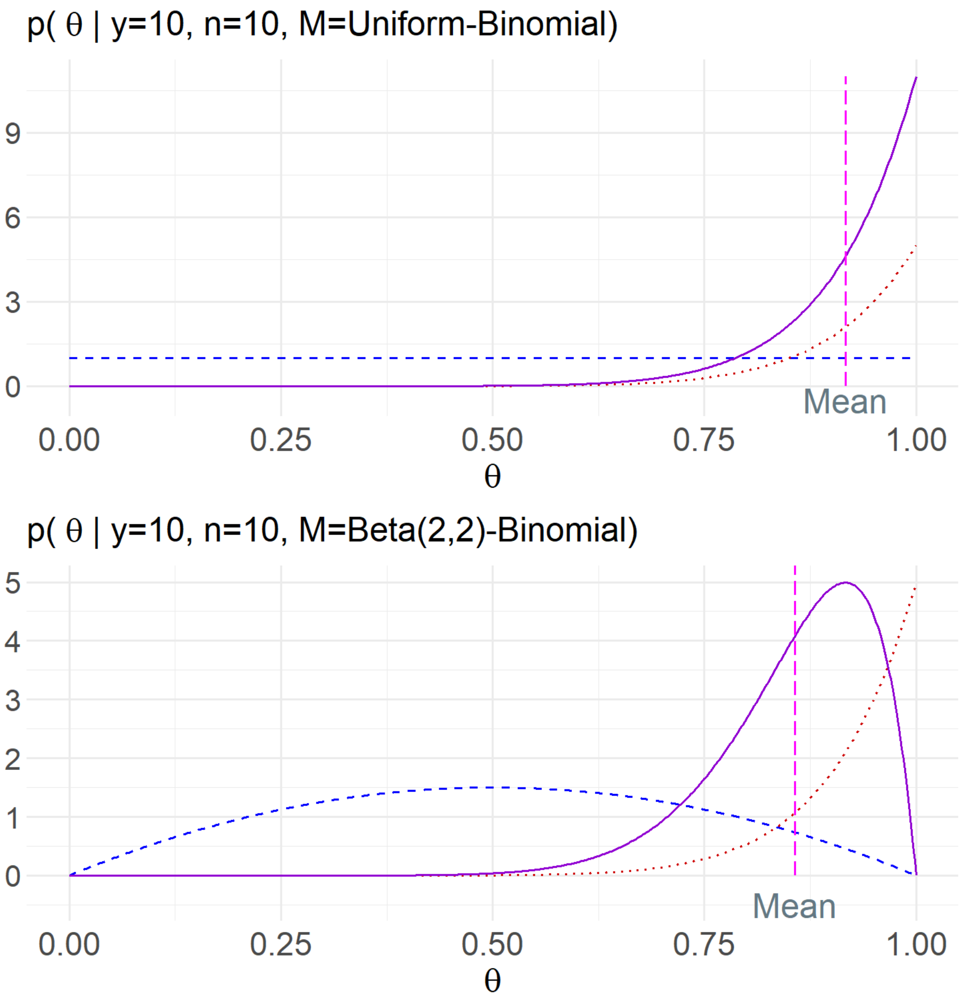


- Comparison of posterior distributions with different parameter values for the Beta prior distribution: Beta priors centered in the population mean, 0.485, and with increasing strenght



Still on Perfect Response Pattern example

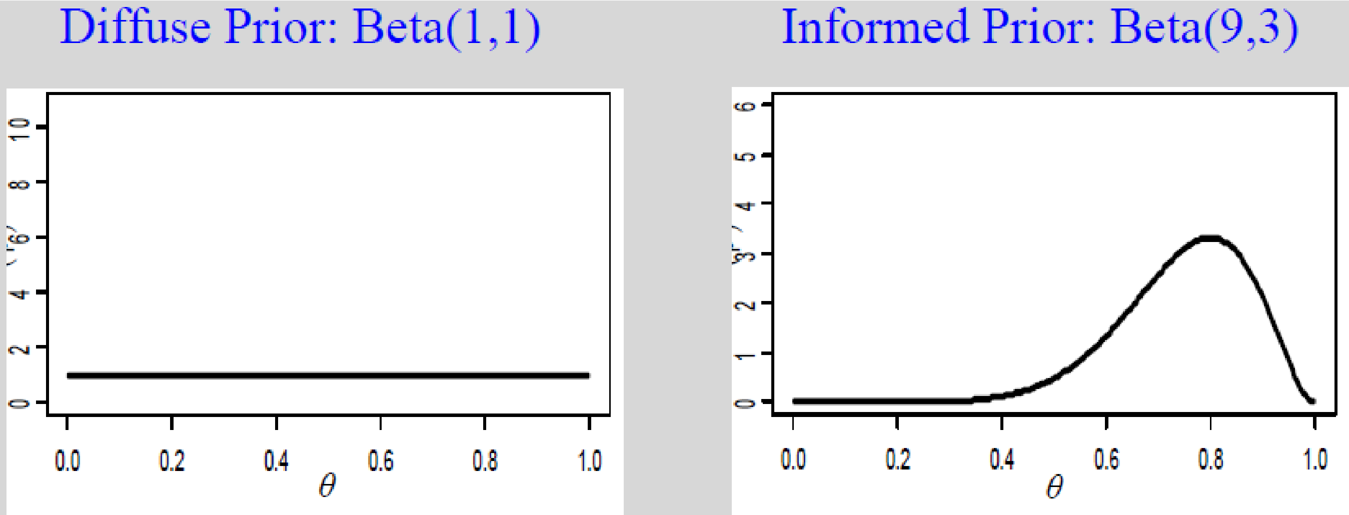
Example: $n = 10, y = 10$ - uniform priori vs Beta(2,2)



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?



Diffuse Prior: Beta(1,1)

Informed Prior: Beta(9,3)

Esempio: *Perfect Response Patterns*

Prior mean = .75

Prior mode = .80

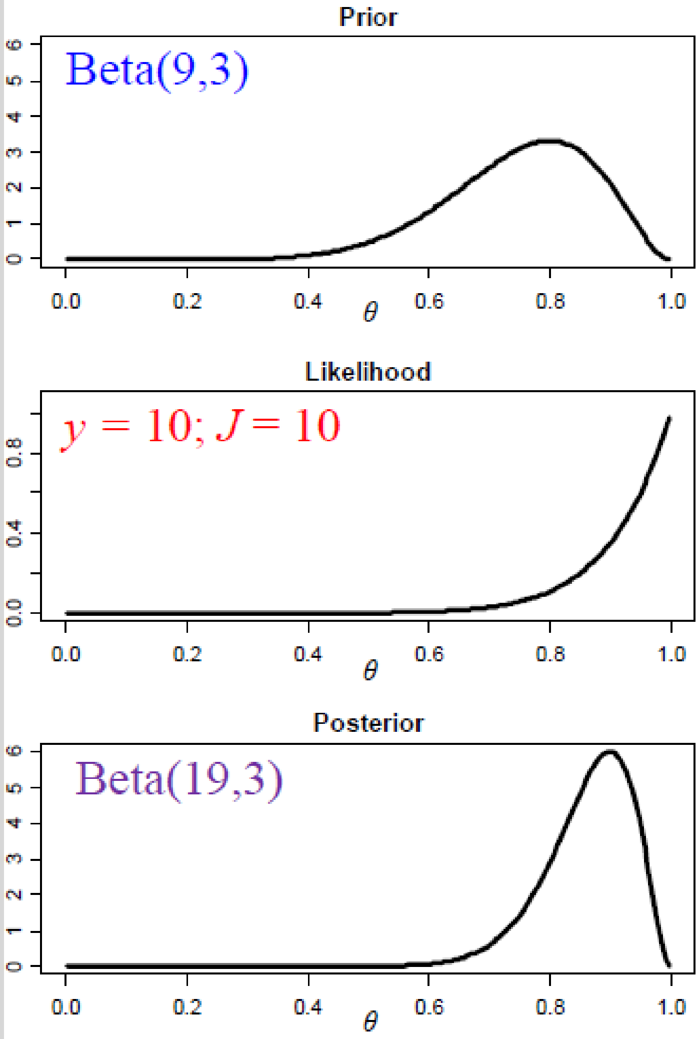
Prior sd = .12

Max. likelihood = 1

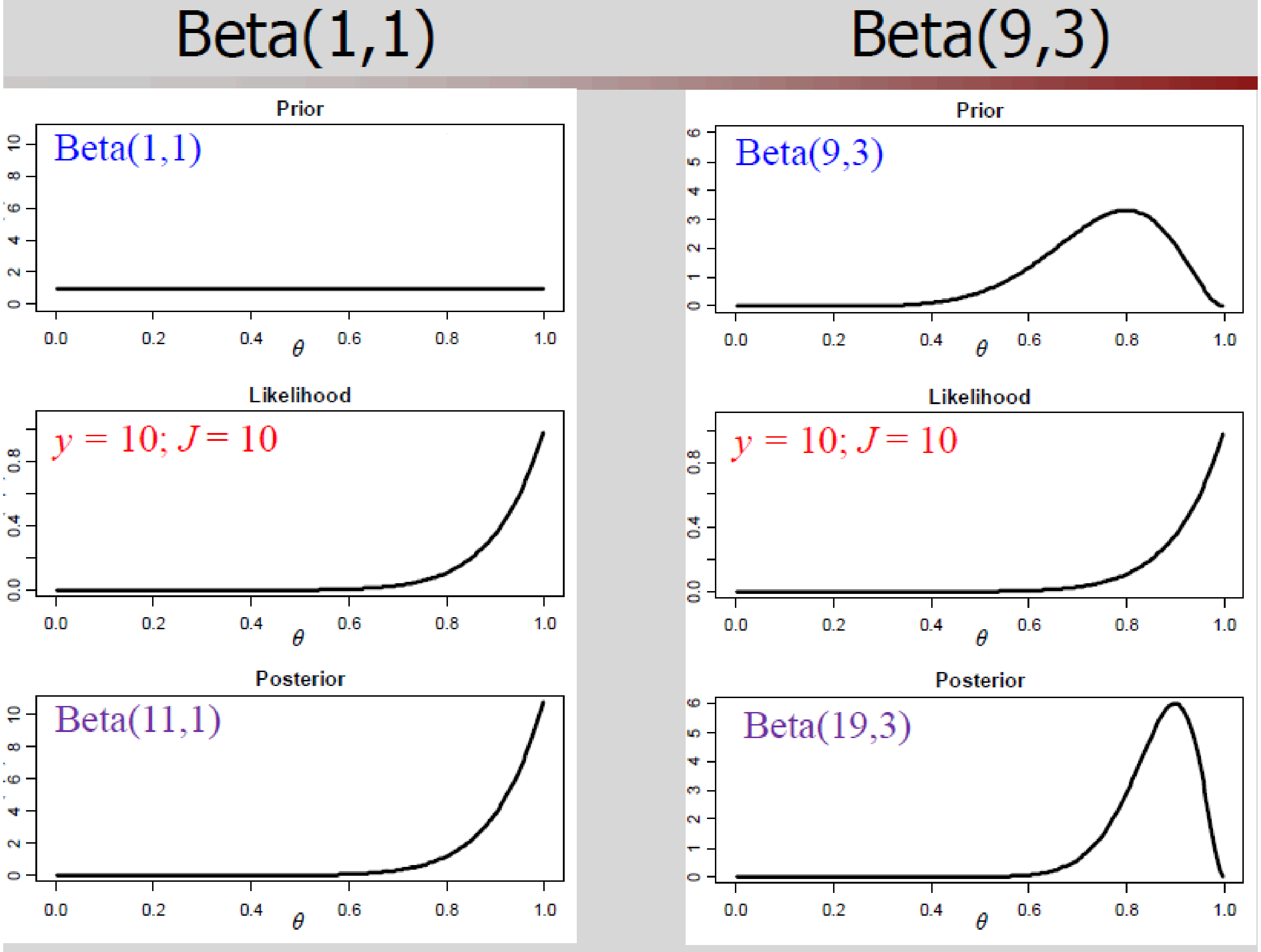
Posterior mean = .86

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

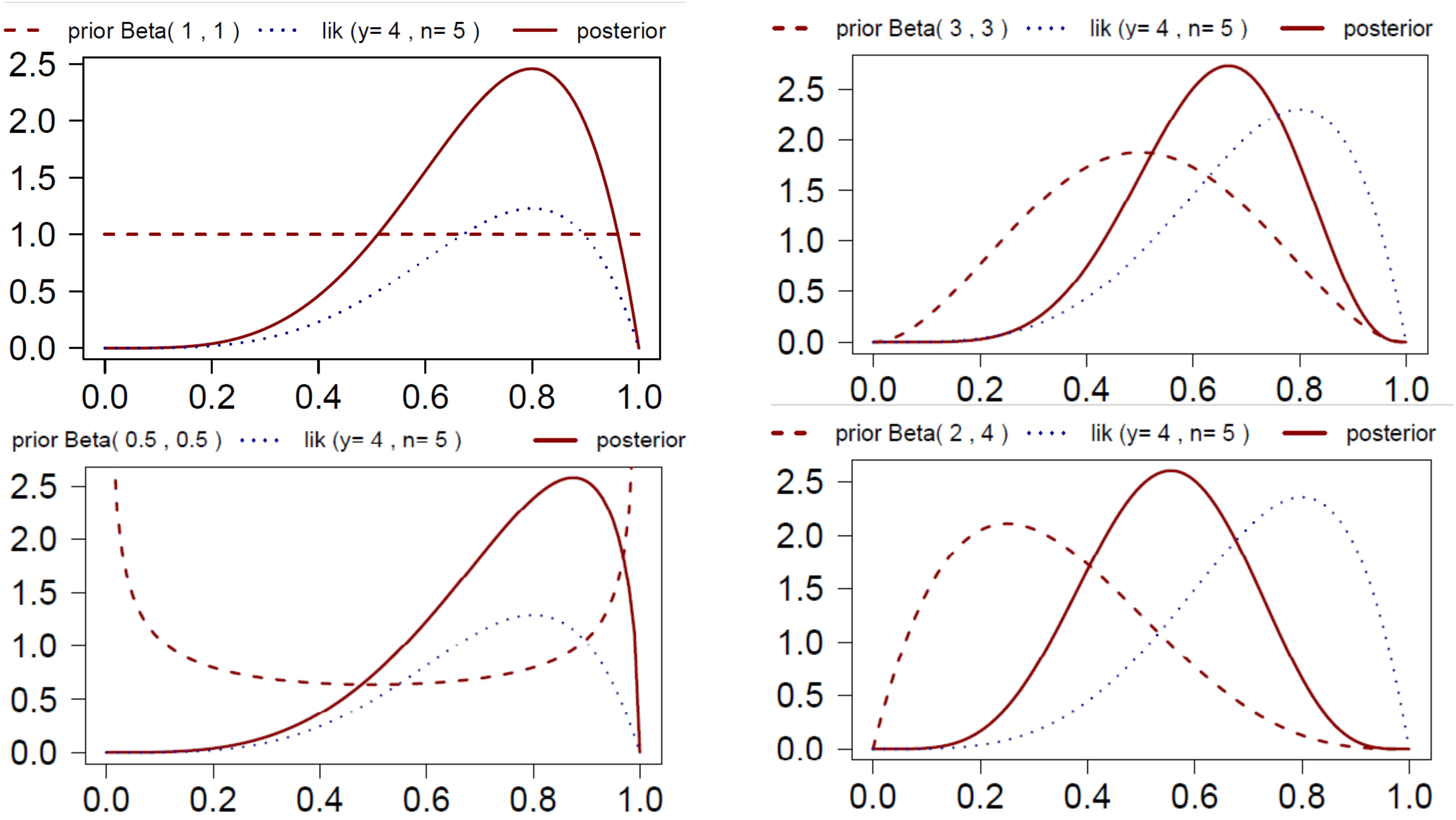
The posterior variance is

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

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

Different priors ⇒ different posteriors

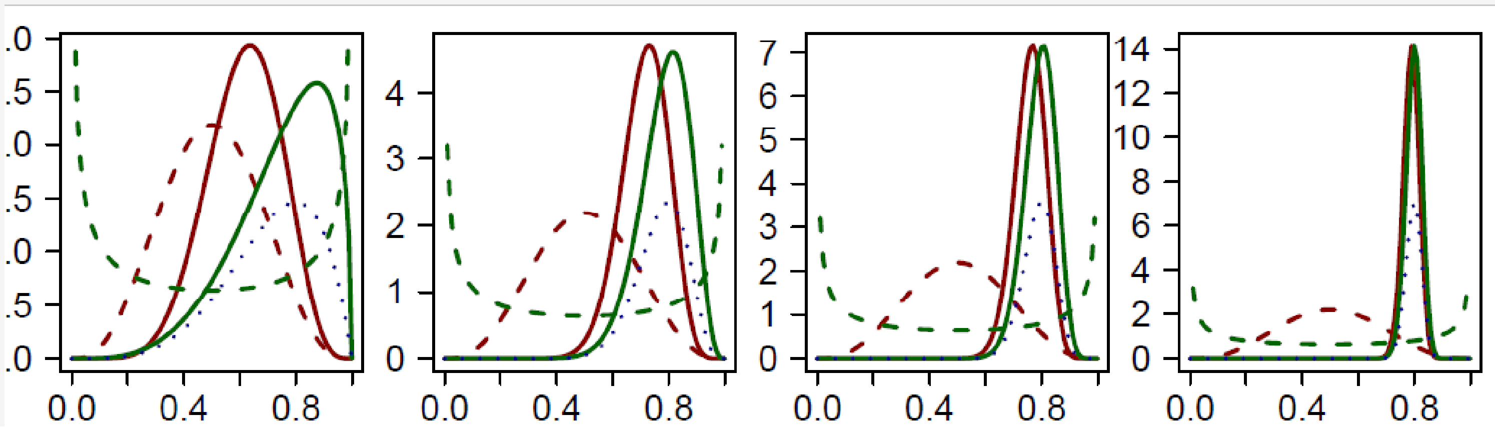
uniform, centered at 0.5, 0-1, rightly asymmetric



Prior effect as n increases

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

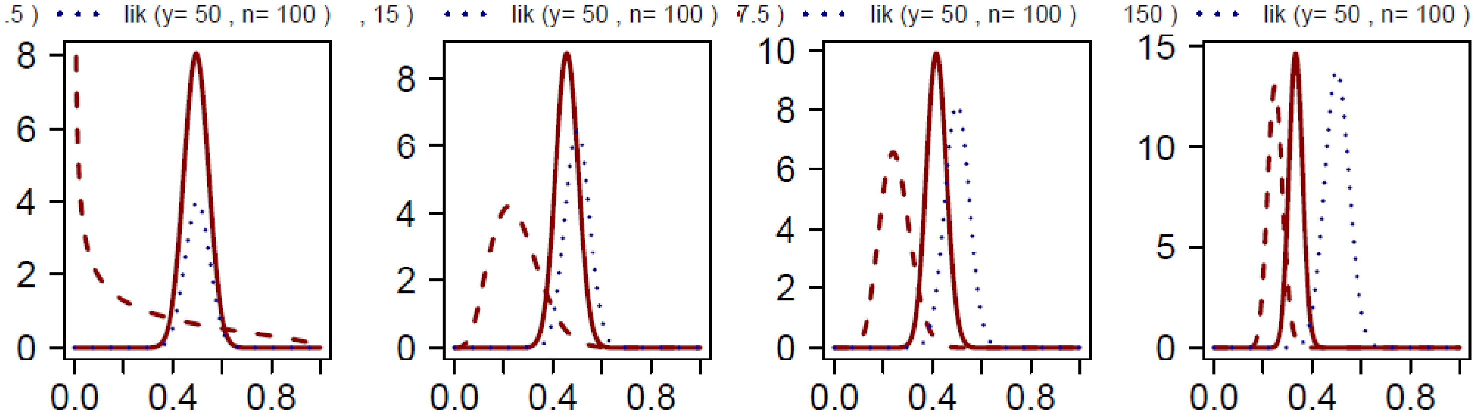
In the following we observe the effect on the posterior of two distinct priors on samples of $n = 5, 20, 50, 200$, always with $y/n = 0.8$



Prior effect as $\alpha + \beta$ increases

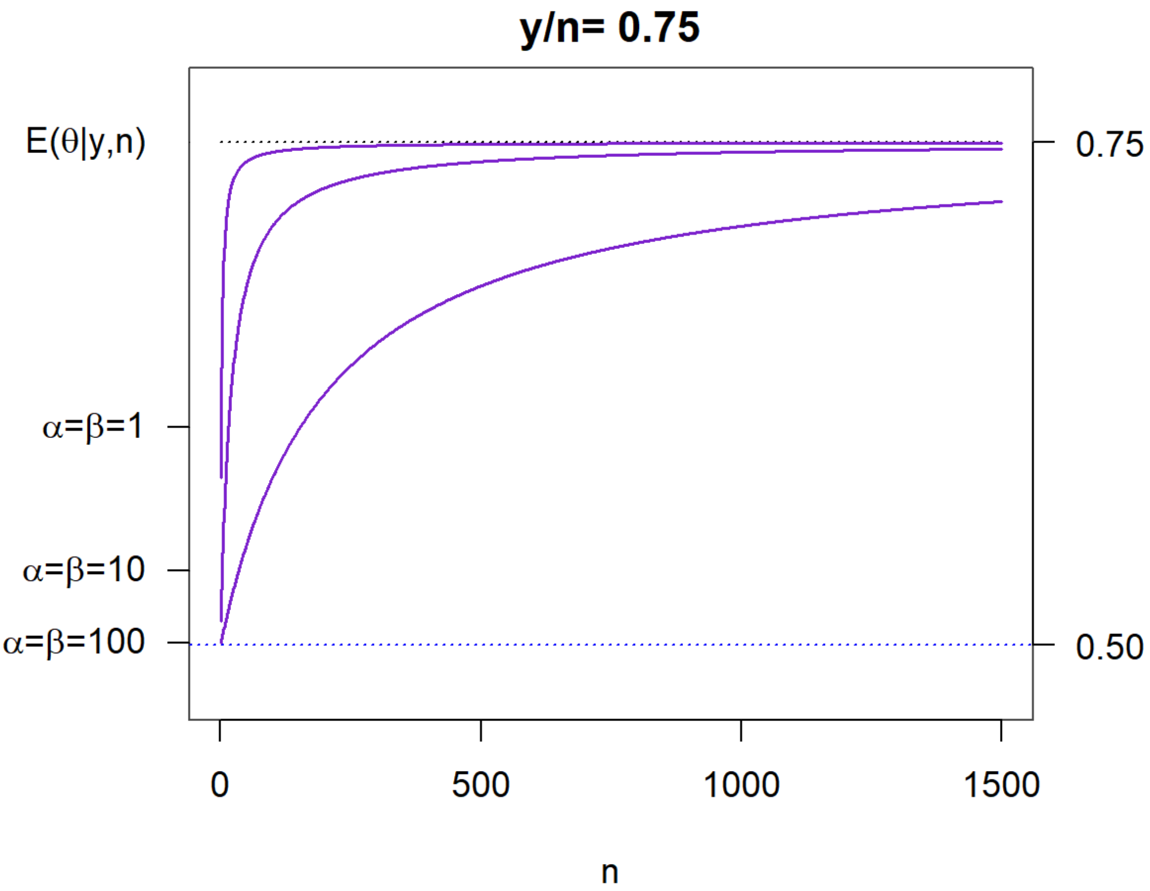
We can see things from another point of view and consider different priors with the same sample.

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



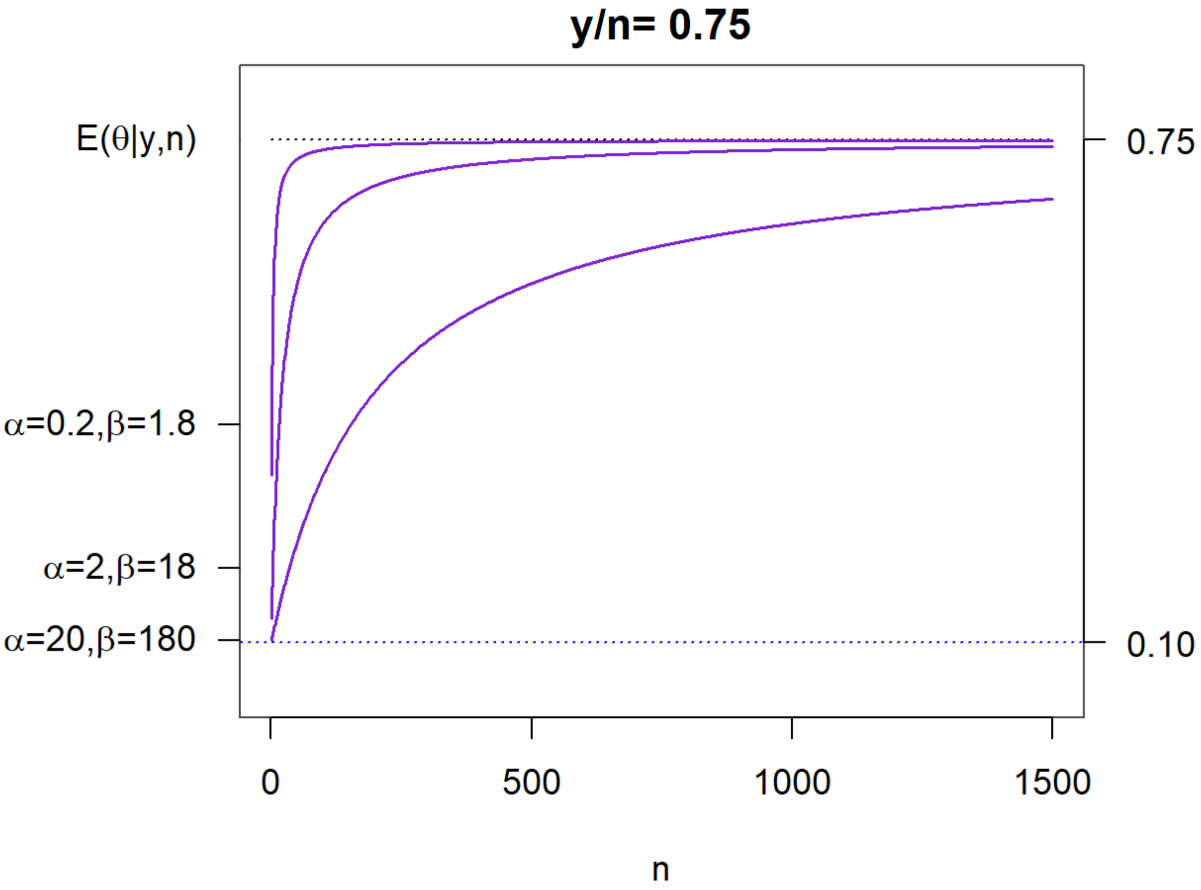
Posterior mean as a function of sample size

Posterior mean as n increases for different priors



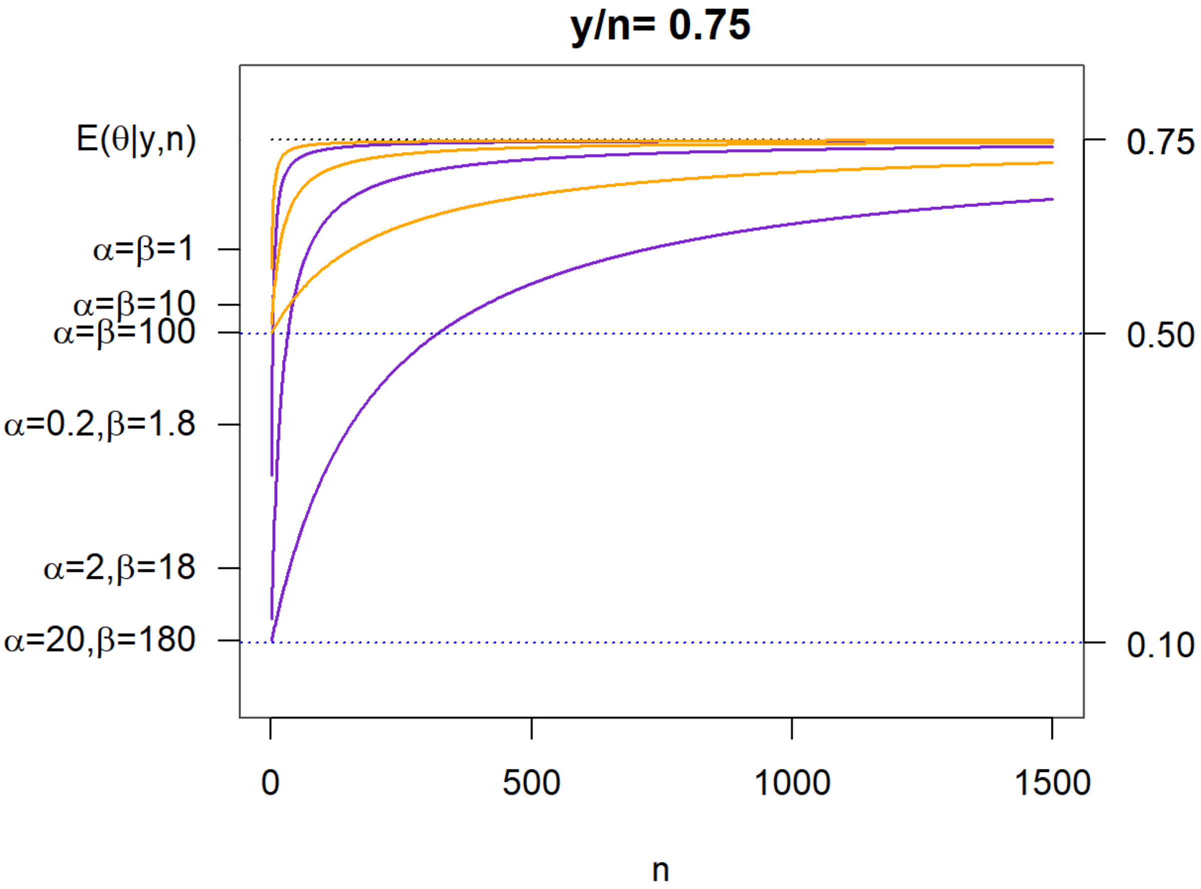
Posterior mean: conflicting priors

Posterior mean as n increases for different priors



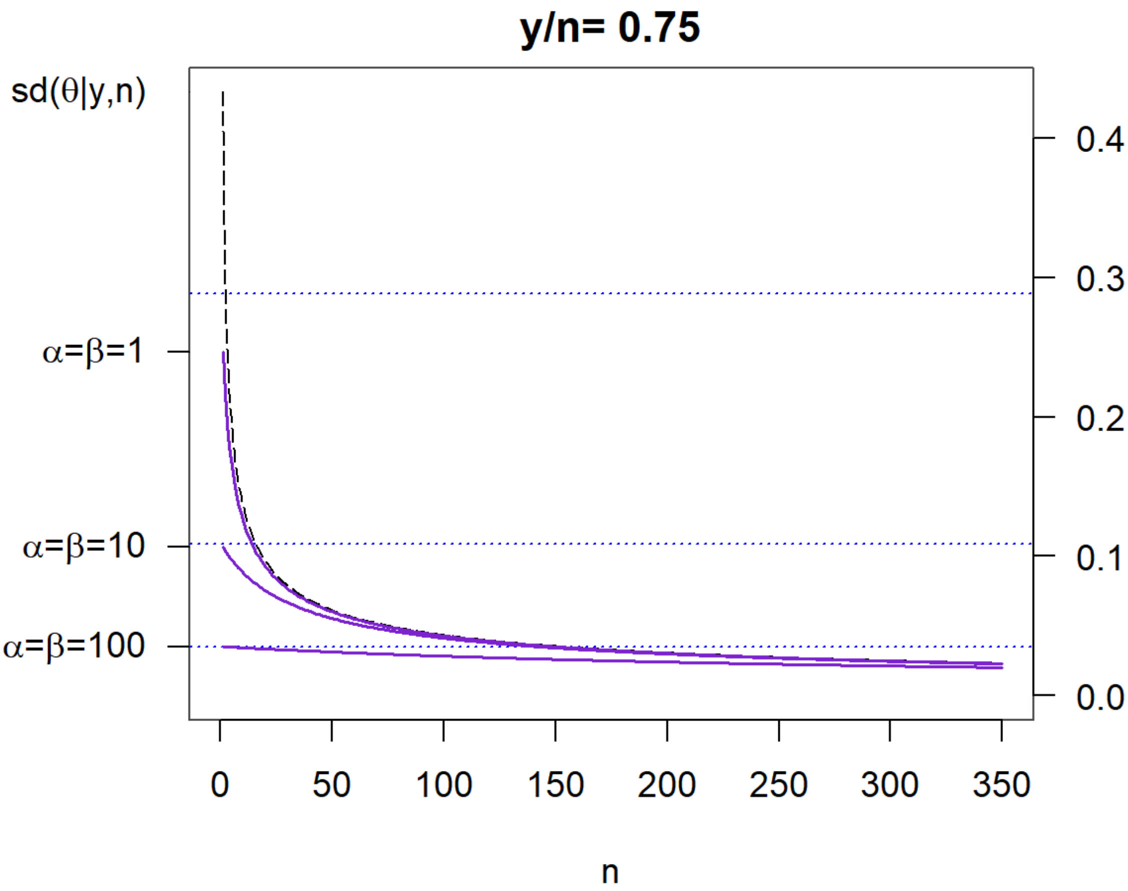
Posterior mean: all together

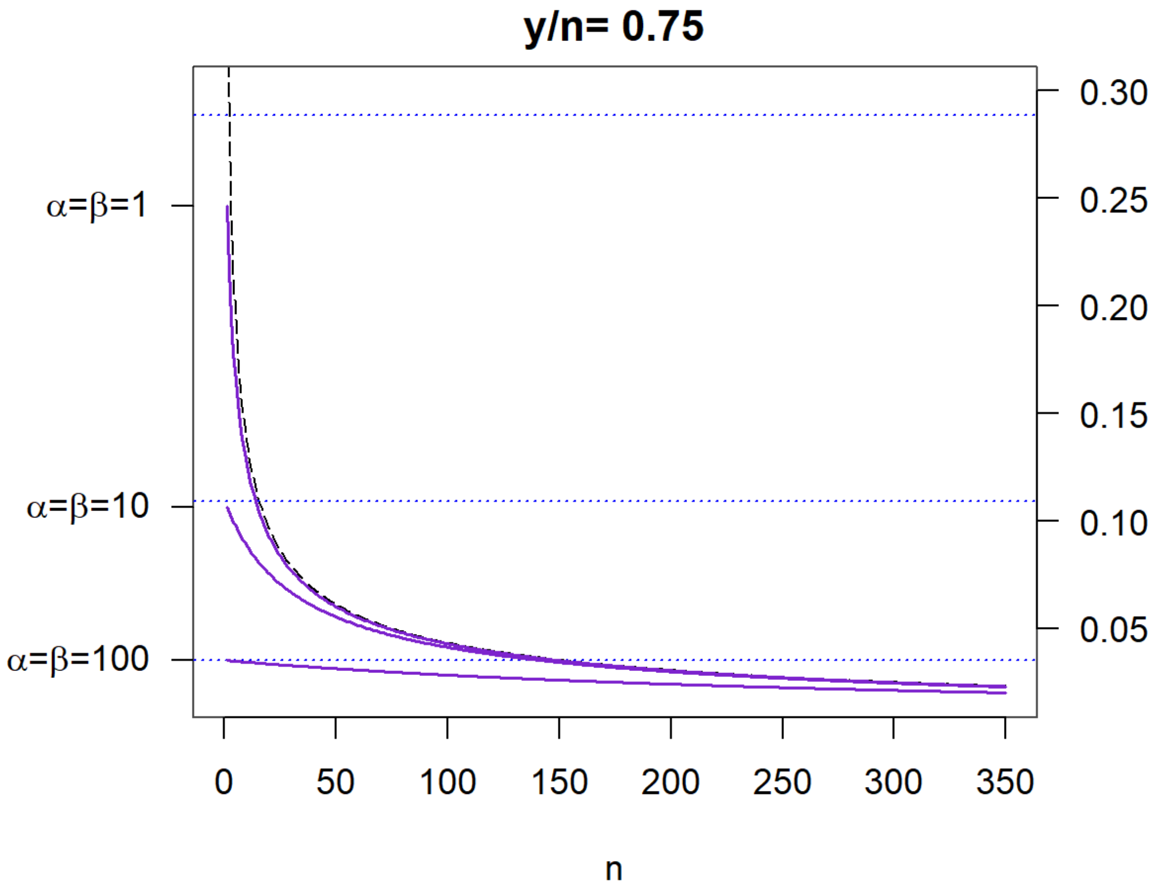
Posterior mean as n increases for different priors



Posterior variance as a function of sample size

Posterior variance as n increases for different priors

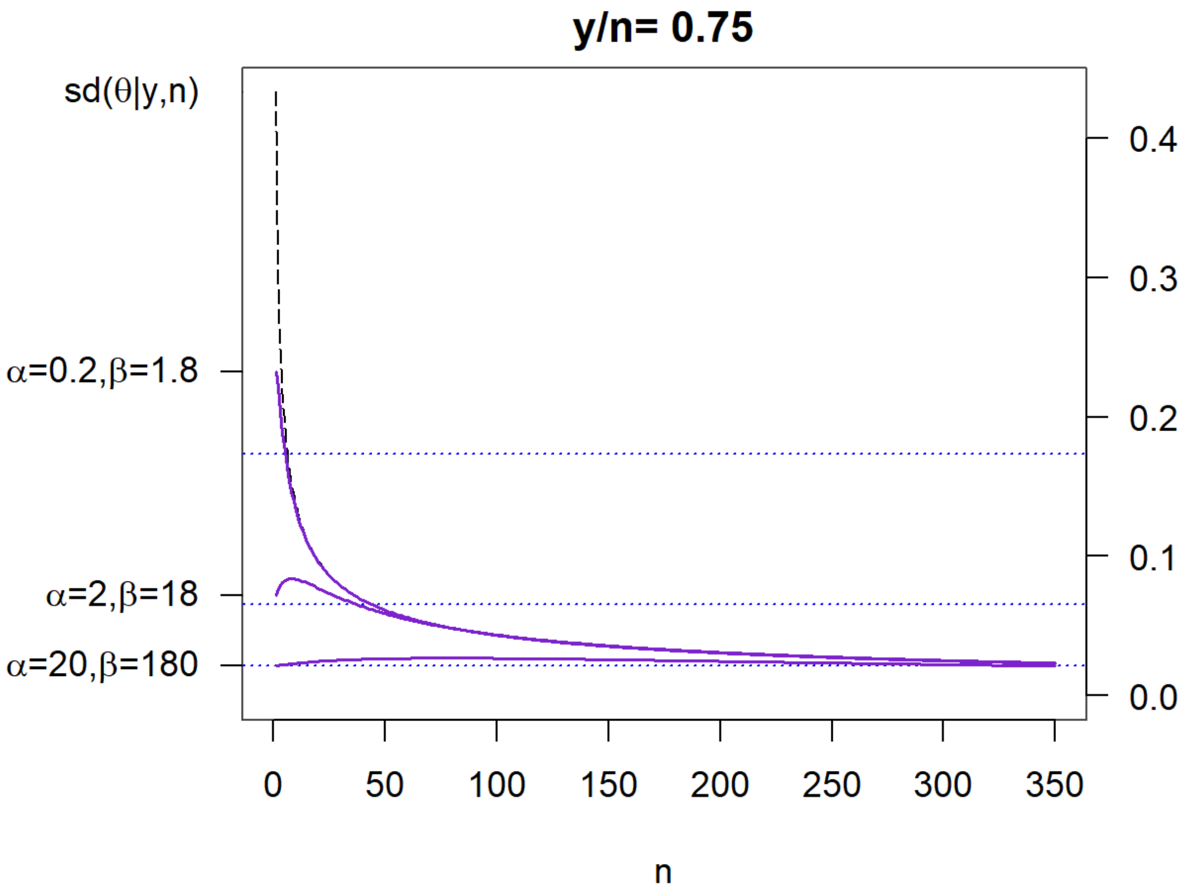


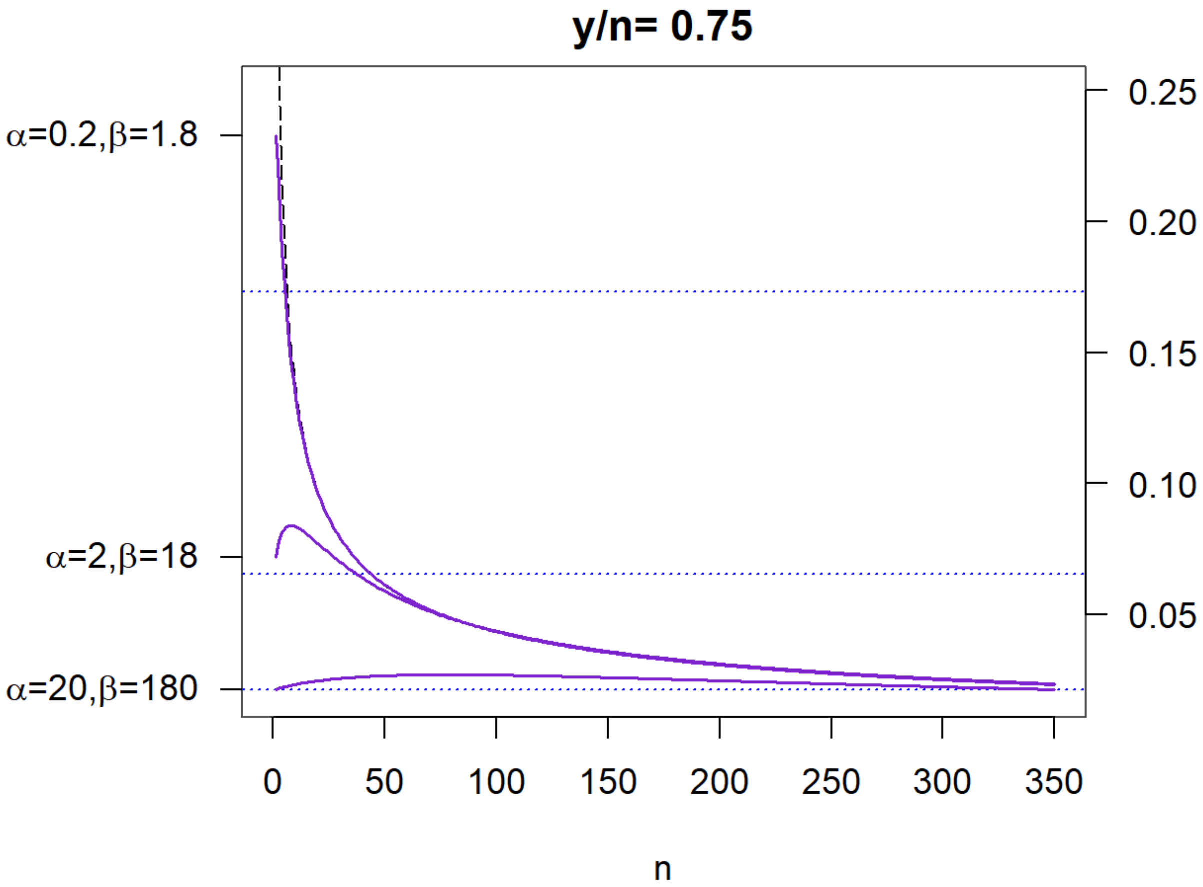


```
## sd(MLE) with n=1:3,10,50,100,350  0.43 0.31 0.25 0.14 0.06 0.04 0.02
## prior sd: 0.289 0.109 0.035

## 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
```

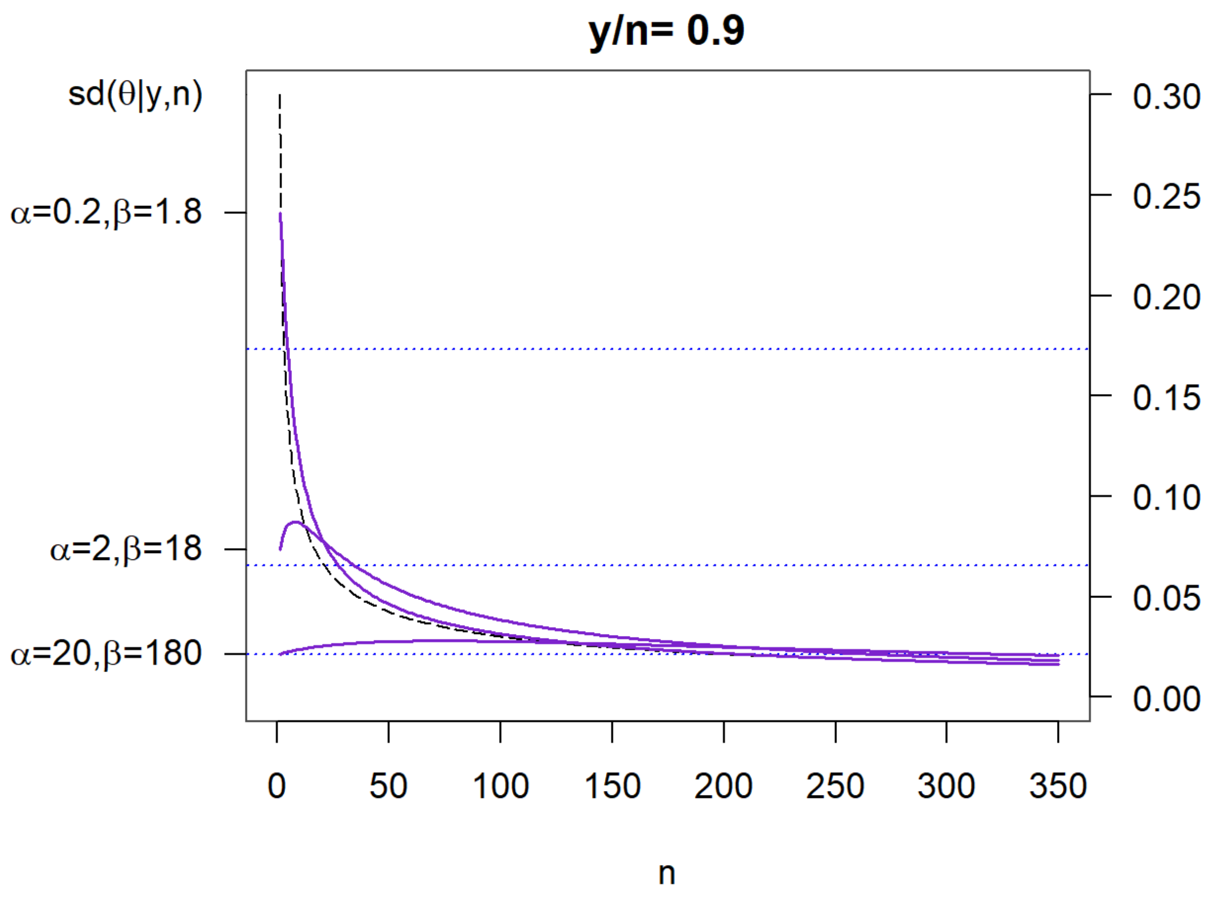
Posterior variance: conflicting priors

Posterior variance as n increases for different priors

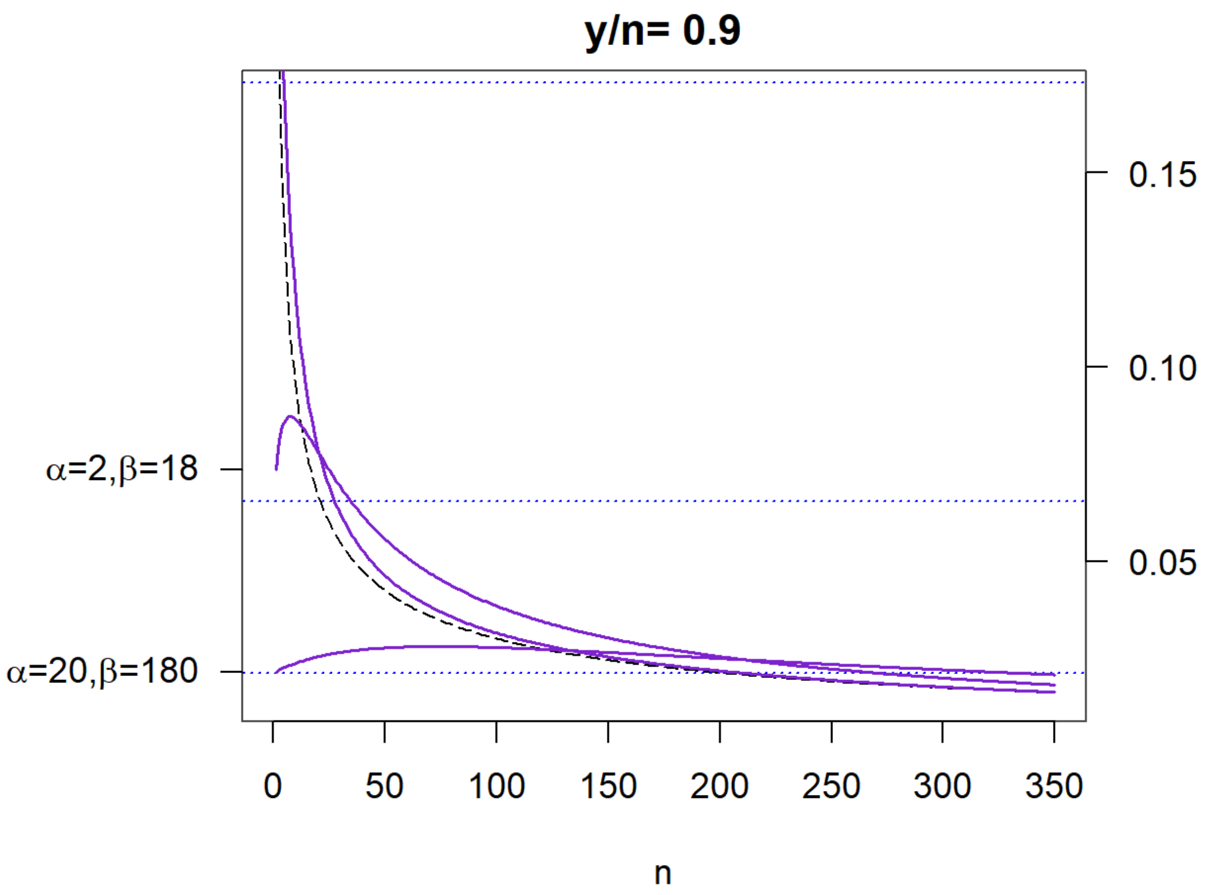




```
## sd(MLE), n=1:3,10,50,100,350  0.433 0.306 0.25 0.137 0.061 0.043 0.023
## prior sd: 0.173 0.065 0.021

## 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
```

Non-informative priors

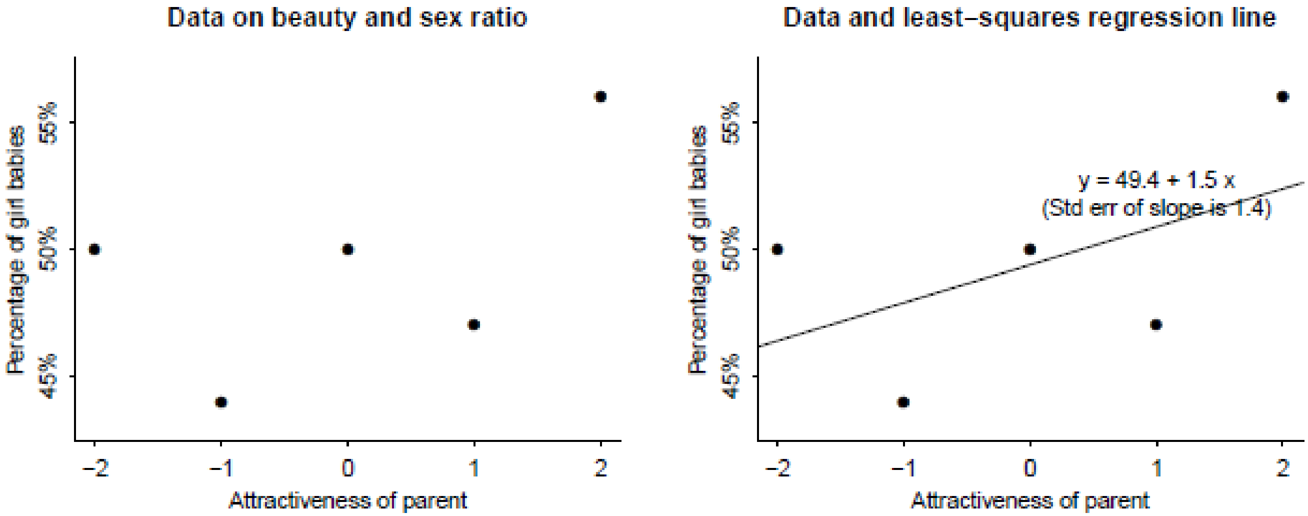
- Vague, flat, diffuse, or **noninformative**
    - tend to "let the data speak for themselves"
    - Noninformative distributions are typically defined as being flat over the entire real axis (e.g., $\propto 1$). Common noninformative priors include wide uniform distributions (e.g., $U(-1000, 1000)$ or $U(0, 1000)$ for positive-only parameters) or diffuse normal distributions (e.g., $N(0, 10000)$)
    - flat is not always true to be noninformative
        - Assigning flat prior distributions to transformed parameters often yields highly skewed, strongly informative priors for the parameter in the original scale.
    - flat can be a stupid choice
        - A more accurate definition of noninformative priors would be 'distributions that possess a range of uncertainty larger than any plausible parameter value'
    - Making the prior flat somewhere can make it non-flat somewhere else
- proper prior have $\int p(\theta) = 1$
- density of improper prior does not have a finite integral
    - posterior sometimes can be still proper

Weakly informative priors

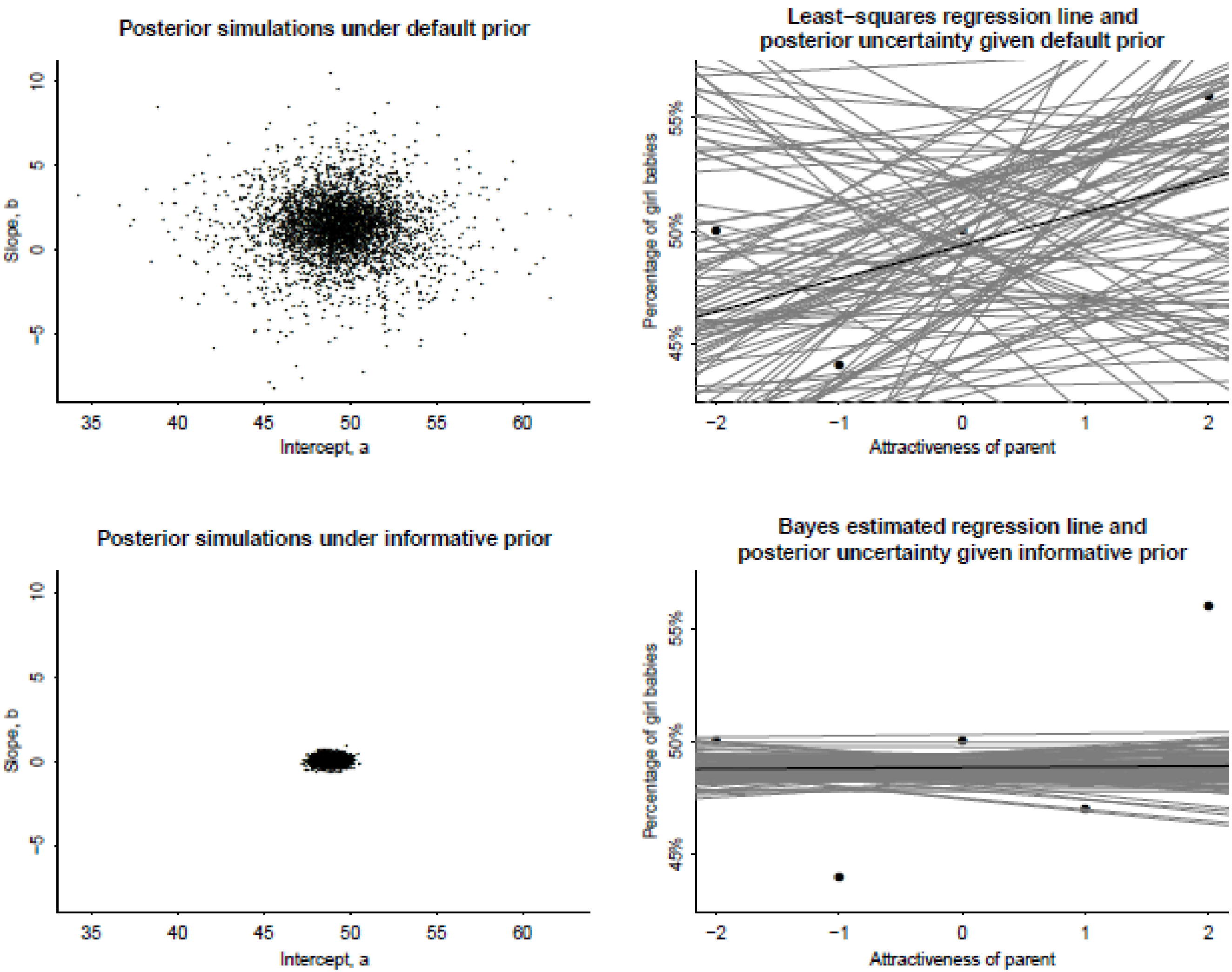
- ***Weakly informative*** produce a better behavior of the posterior from a computational point of view
  - quite often there is at least some knowledge of the scale
  - also useful if you have information from previous observations, but you are not sure how applicable that information is to the uncertainty of the new case
- Construction
  - Start with some version of a non-informative prior and then **add** enough information to make the inferences "reasonable".
  - Start with a strong, very informative prior and **expand** it to account for uncertainty in your prior beliefs and in the applicability of any prior based on past history to new data.
- Stan team's preliminary choice recommendations https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

Example of informative prior

- The percentage of female births is remarkably stable at around 48.5% , rarely varying by more than 0.5% from this rate
- There is a study on the percentage of female births among parents in attractiveness categories 1–5 (rated by interviewers in a face-to-face survey)



Example of informative prior



Posterior distribution estimation via simulation

Analytical solutions (doing mathematical calculations)

- Facilitated by conjugate priors
    - For $x \sim \mathrm{Binom}(\theta, n), \theta \sim \mathrm{Beta} \rightarrow \theta|x \sim \mathrm{Beta}$

  When we cannot obtain the posterior analytically, it is necessary to estimate or approximate it in some way

- It can be approximated or estimated
- A flexible and general approach to estimating distributions is necessary
    - Realized by simulation
    - A little now, more later

Simulation-based Estimation

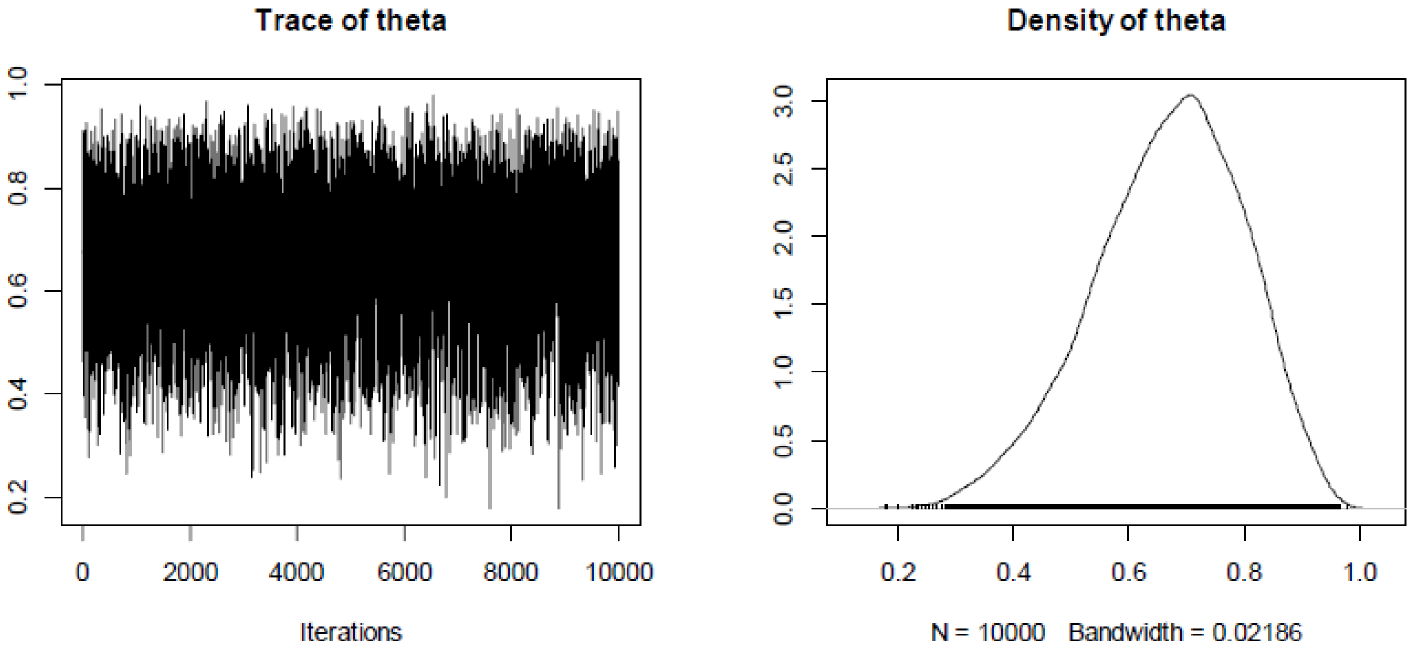
A sampling algorithm is constructed to ***simulate*** or ***draw from the*** posterior. Many such draws are sampled, which serve to empirically approximate the posterior distribution, and can be used to empirically approximate the summary statistics.

Monte Carlo Principle:

> 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)

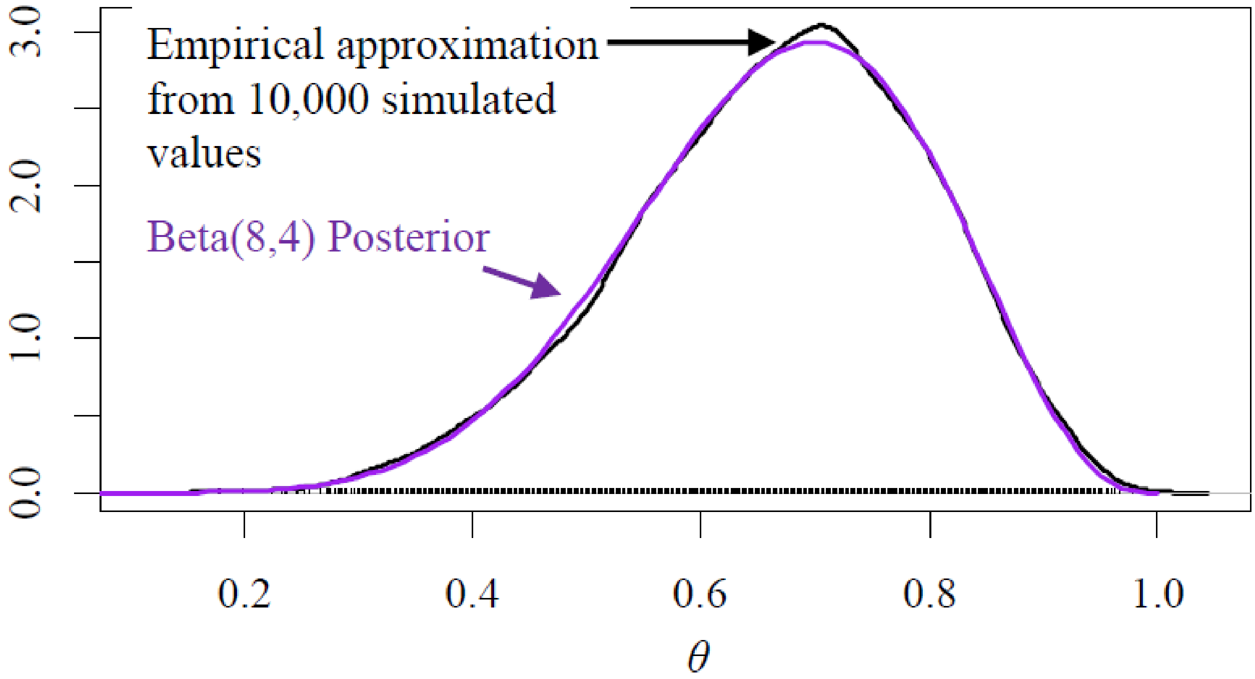
Simulation of the posterior in the Beta-Binomial model

A sampling algorithm is constructed to ***simulate*** or ***extract from the*** posterior.



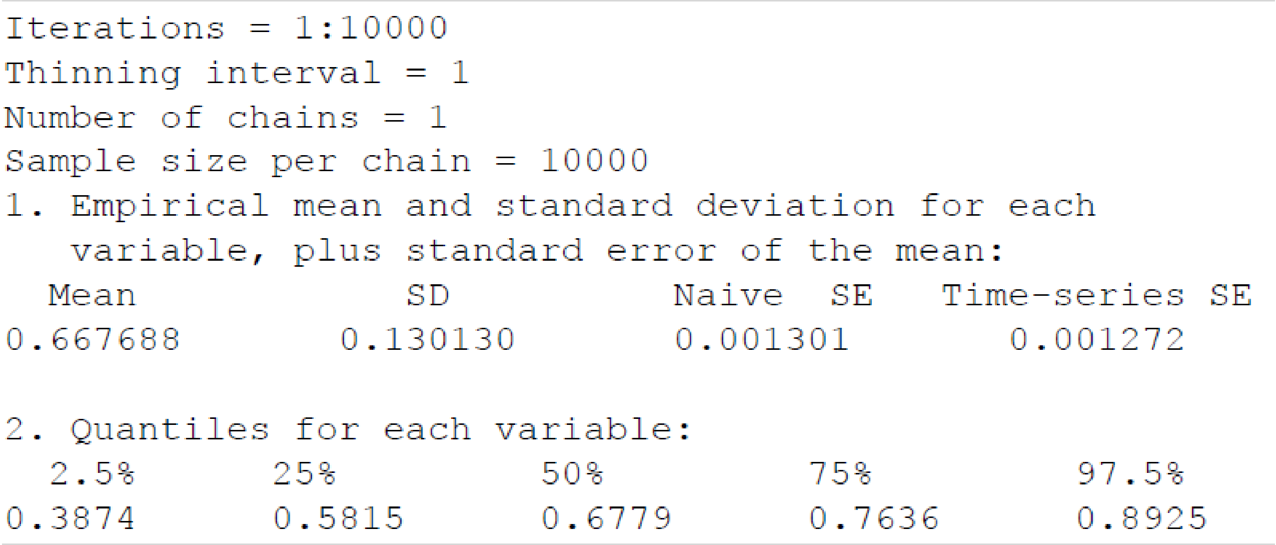
Beta-Binomial Model: Density

A lot of these extractions are collected, which serve to empirically approximate the posterior distribution



Beta-Binomial Model: Summary Statistics

and also to empirically approximate the summary statistics.



```
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
```

Marginal Notes

Model and likelihood

Term $p(y|\theta, M)$ has two different names depending on the situation. Due to the short notation used, there is possibility of confusion.

1. Term $p(y|\theta, M)$ is called a **model** (sometimes more specifically *observation model* or *statistical model*) when it is used to describe uncertainty about y given θ e M. Longer notation $p_y(y|\theta, M)$ shows explicitly that it is a function of y.

2. In Bayes rule, the term $p(y|\theta, M)$ is called **likelihood function**. Posterior distribution describes the probability (or probability density) for different values of θ given a fixed y, and thus when the posterior is computed the terms on the right hand side (in Bayes rule) are also evaluated as a function of θ given a fixed y. Longer notation $p_\theta(y|\theta, M)$ shows explicitly that it is a function of θ.
   Term has it's own name (likelihood) to make the difference to the model. The likelihood function is unnormalized probability distribution describing uncertainty related to θ (and that's why Bayes rule has the normalization term to get the posterior distribution).

Ambiguous notation in statistics

In $p(y|\theta)$

- y can be variable or value
  - we could clarify by using $p(Y|\theta)$ or $p(y|\theta)$
- θ can be variable or value
  - we could clarify by using $p(y|\Theta)$ o $p(y|\theta)$
- p can be a discrete or continuous function of y or θ
  - we could clarify by using P_Y, P_Θ, p_Y or p_Θ
- $P_Y(Y|\Theta = \theta)$ is a probability mass function, sampling distribution, observation model
- $P(Y = y|\Theta = \theta)$ is a probability
- $P_\Theta(Y = y|\Theta)$ is a likelihood function (can be discrete or continuous)
- $p_Y(Y|\Theta = \theta)$ is a probability density function, sampling distribution, observation model
- $p(Y = y|\Theta = \theta)$ is a density
- $p_\Theta(Y = y|\Theta)$ is a likelihood function (can be discrete or continuous)
- y and θcan also be mix of continuous and discrete

Back to 3-step general approach