Introduction to Bayesian inference

Master's Degree in Biostatistics

Joaquín Martínez-Minaya, 2021-11-10
Universitat Politècnica de València
https://smeg-bayes.org/
http://vabar.es/
jmarmin@upv.es

Outline

1. What is Bayesian statistics?

2. Doing inference on the parameters

3. Predictions

4. Hierarchical Bayesian models

1. What is Bayesian statistics?

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Basics of Bayesian Statistics

- Bayesian approach is another **way of understanding and doing Statistics**.
- It is **NOT another field** inside Statistics (like spatial statistics or time series analysis).
- Therefore, when using Statistics to solve a problem (to make inference and prediction about the unknown parameters of the proposed model) we could do it using the **frequentist or Bayesian approach**.
- The most known approach is the **frequentist**: a.k.a. classical (as it is the usual way to perform estimation, hypothesis testing and prediction of the parameters or each function of them).

Fundamentals of the Bayesian approach

Reasoning in terms of probability

- about the observed in the sampling: data
- … but ALSO about the unknown and unobserved: parameters
- **UNCERTAINTY** $\equiv$ **PROBABILITY**

We express our knowledge about something through probability distributions.

- **Likelihood** from the model assumed for the data, $l(\theta; x)$.
- Before observing data: **Prior distribution** about the parameters $p(\theta)$.
- Use **Bayes theorem** to update information about the parameters using data observed.

2. Doing inference on the parameters

Bayesian inference

- Construction of the **joint distribution** about the unknown elements of the problem:
  - $l(\theta; x) = p(\boldsymbol{x}|\boldsymbol{\theta})$ is the **likelihood** function of the observed data,
  - $p(\boldsymbol{\theta})$ is the **prior distribution**,

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})\,.$$

- Using **Bayes theorem** to obtain the posterior distribution:

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{\theta})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}$$

- As $p(\boldsymbol{x})$ does not depend on $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto p(\boldsymbol{\theta}) \times p(\boldsymbol{x}|\boldsymbol{\theta})\,.$$

Example

- Liga Santander is one of the famous league around the world. In this example, we use data of the last 10 seasons in order to know the chance of success (π) to score a penalty for **Valencia Club de Fútbol**.
  - **Likelihood**

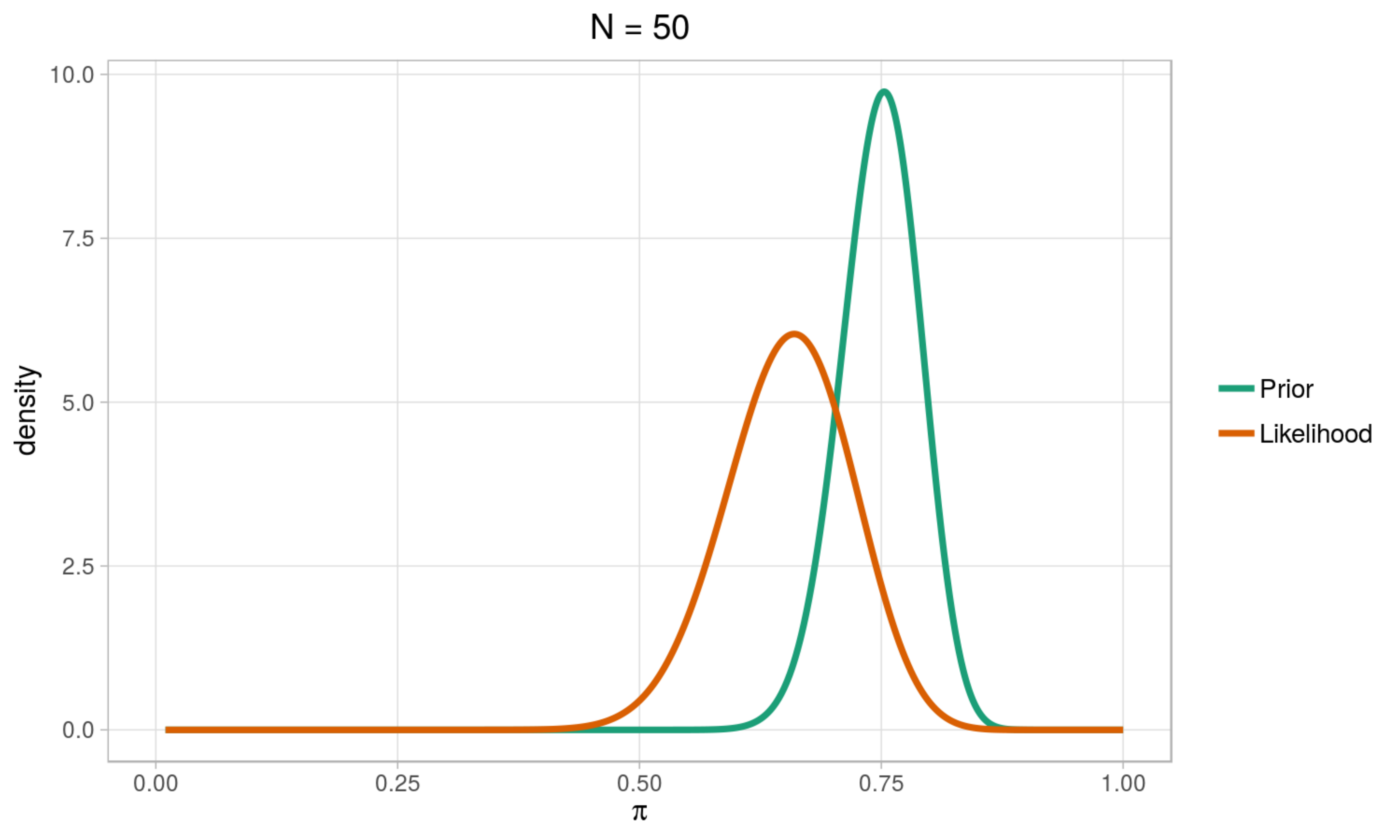
$$p(\boldsymbol{y}|\pi) = \ell(\pi) = \pi^k (1-\pi)^{n-k}$$

  k: times that a player score a penalty.

- After asking some experts, we end up with a 75 percentage chance to score through a penalty. We use a **conjugate prior distribution**, the Beta distribution, to express this information. The corresponding values for a and b are with parameters $a = 83.46$ and $b = 28.05$.
  - **Prior distribution**

$$p(\pi) \propto \pi^{a-1}(1-\pi)^{b-1}$$

Example. Likelihood *vs* Prior



Posterior distribution. Bayesian learning process

Estimating the probability to score

- **Likelihood**

$$p(\boldsymbol{y}|\pi) = \ell(\pi) = \pi^k(1-\pi)^{1-k}$$

- **Prior distribution**

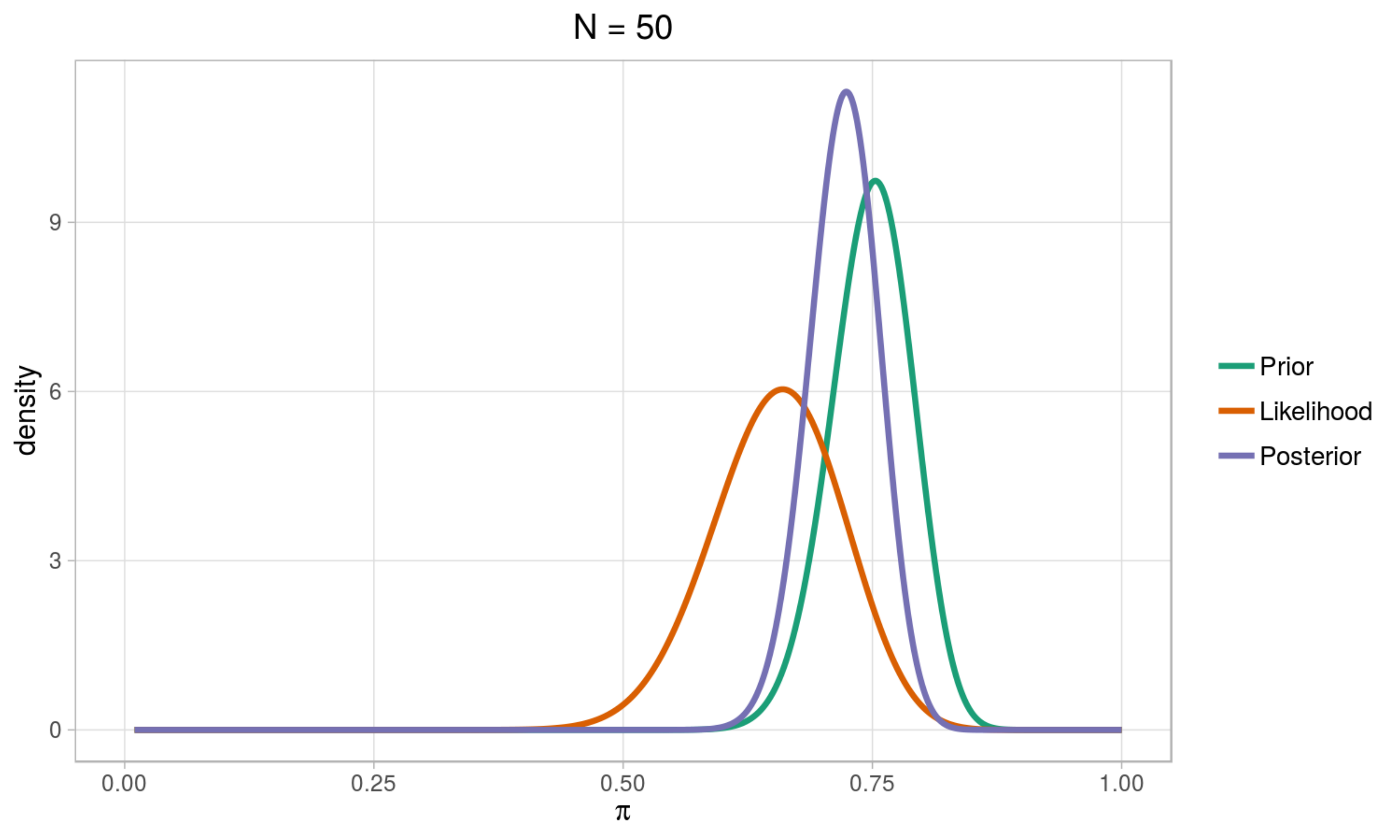
$$p(\pi) \propto \pi^{a-1}(1-\pi)^{b-1}$$

- **Posterior distribution**

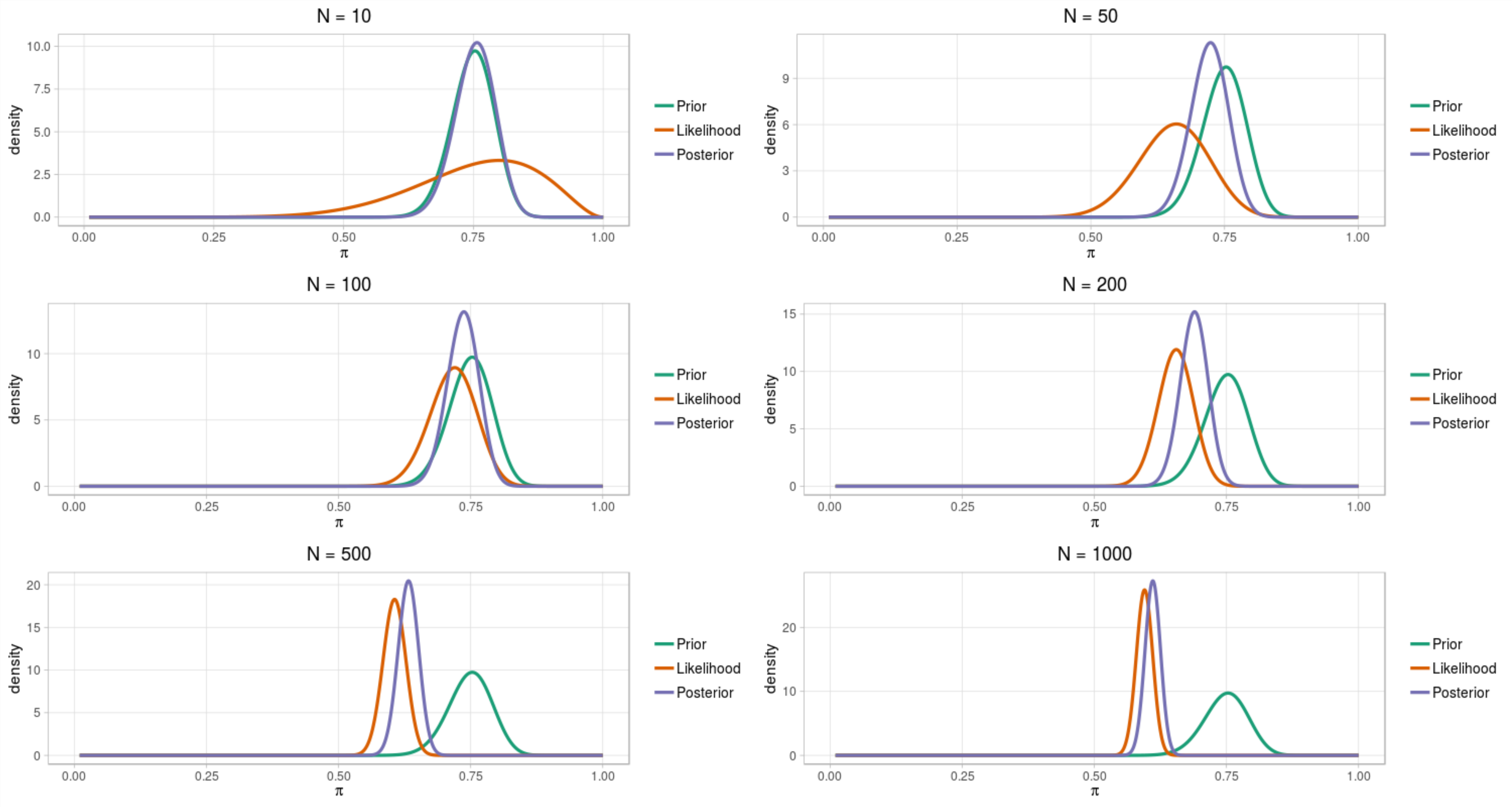
$$p(\pi|\boldsymbol{y}) \propto p(\boldsymbol{x}|\pi) \times p(\pi) \propto \pi^{k+a-1}(1-\pi)^{n-k+b-1}$$

$$\pi|\boldsymbol{x} \sim \mathrm{Beta}(k+a, n-k+b)$$

Posterior distribution



Data vs prior information



3. Predictions

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Predictions

Prior predictive distribution

- Using just the **previous information** about the population.
- **Before performing the experiment** it is possible to infer about the most and least probable values to be observed.

$$p(y_{pred}) = \int p(y_{pred} \mid \theta) p(\theta) d\theta$$

Posterior predictive distribution

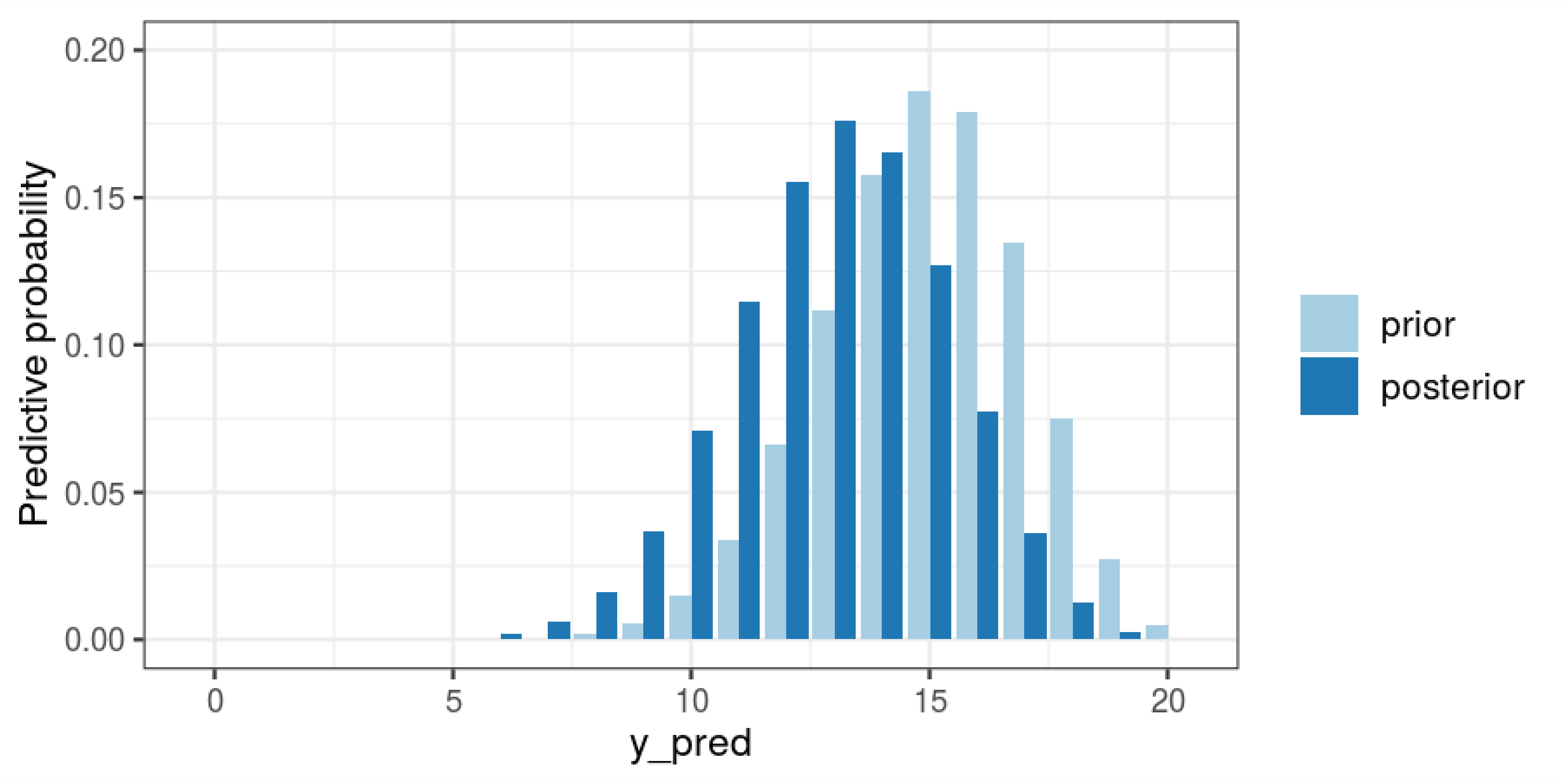
- Using the **updated** information after performing the experiment.
- Allows us to infer about the most and least probable values to be observed if we would **repeat the experiment** in the future (in the same conditions).

Prior *vs* Posterior predictive

- José Bordalás wants to know if he can trust in their players to score the next penalty. In this figure we see what happened if we take into account just the expert knowledge or the expert knowledge and the data.

Prior *vs* Posterior predictive

- José Bordalás want to know if he can trust in their team to score the next 20 penalties. How many of then we predict that they are going to score?



But...

- Each team have many **players**

- In the league there are different **teams**

- En each country there is a different **league**

- In addition to the league, there exist other **competitions**: Champions, Europa league

- There exist a hierarchy

How can we model that?

4. Hierarchical Bayesian models

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Again we talk about football

- We consider same experiment in **10 different teams**
- How can we model this situation? and what can we conclude?
- More generally, how can we incorporate **random effects**?

Three ways to do so

1. Consider that **all teams have the same characteristics**.

- Apply a **joint analysis** to all the teams.
- The probability of score a penalty (π) is the **same in all teams**.
- Observations are independent and identically distributed.

$$y_i \mid \pi \sim \mathrm{Ber}(\pi)$$

$$\pi \sim \mathrm{Beta}(a, b)\ ,\ \text{with a and b fixed}$$

- Usual **estimator**, m.l.e.:

$$\hat{\pi} = \frac{\sum_i r_i}{\sum_i n_i}$$

being r_i the number of penalties scored and n_i the total number of penalties.

Three ways to do so

2. Consider that **each team is different and has nonthing in common with the others**.

- Apply an analysis to **each team separately**.
- Assume a **different proportion of presence** in each one: π_j, $j = 1, \ldots, 10$.
- Observations are independent but are **distributed differently in each team**.
- **Likelihood** is different for each team. For each j,

$$y_{ij} \mid \pi_j \sim \mathrm{Ber}(\pi_j)$$

$$\pi_j \sim \mathrm{Beta}(a_j, b_j)\ , \text{ with } a_j \text{ and } b_j \text{ fixed}$$

- The proportion in team j is **estimated** with the sampling information obtained on it:

$$\hat{\pi}_j = \frac{\sum_i r_{ij}}{\sum_i n_{ij}}$$

In view of the two possible modelings

- Is it reasonable to assume **the same proportion of presence** in all teams?

- There are reasons to suggest that **there is variability in those proportions**:

  - The teams do not behave the same way.
  - The observations of the same team are more similar among themselves than when they are from different teams.

- Is it reasonable to think that **there is no relationship between the proportions of presence** of the different teams?

Although not identical, **teams** at least **are similar**.

Three ways to do so

3. Consider a **hierarchical model**.

- The parametric vector $\boldsymbol{\pi} = \pi_1, \ldots, \pi_{10}$ is a **random sampling from a common distribution** that depends on a vector of **hyperparameters**, $\boldsymbol{\psi}$, partial or totally unknown.

- The model

  - **Likelihood**

$$y_{ij} \mid \pi_j \sim \text{Ber}(\pi_j)\ , j = 1, \ldots, 10$$

  - **Random effects**

$$\pi_j \sim \text{Beta}(a, b)$$

  - **Hyperparameters**, $\boldsymbol{\psi} = (a, b)$

$$a \sim p(a),\ b \sim p(b)$$

Numerical approaches

- When applying Bayesian Statistics, most of the usual models do not yield to **analytical expressions for the posterior** neither the posterior predictive distributions.
- Most of the **complications that appear in the Bayesian methodology** come from the resolution of integrals that appear when applying the learning process:
  - The normalization constant of the posterior distribution,
  - moments and quantiles of the posterior,
  - credible regions, probabilities in the contrasts, etc.

Solutions:

- **Monte Carlo methods: MCMC**.
- **INLA**.

The End

UNIVERSITAT POLITÈCNICA DE VALÈNCIA