ADIM: An introduction to Bayesian inference

Master's Degree in Data Analysis, Process Improvement and Decision Support Engineering

Joaquín Martínez-Minaya, 2024-12-02

UNIVERSITA
POLITÈCNIC
DE VALÈNC

VAlencia BAyesian Research Group
Statistical Modeling Ecology Group
Grupo de Ingeniería Estadística Multivariante
jmarmin@eio.upv.es

What do the following events have in common?

UNIVERSITA
POLITÈCNIC
DE VALÈNC

Palomares Bomb

- In a routine **Cold War** operation on 17 January 1966 in Palomares (Almería, Spain), an incident happened.
- At 10:22 a.m. that day, **two aircrafts collided over the skies of Almeria**: a KC-135 from Moron Air Base (Seville) was about to refuel a B-52 from Turkey that was returning to North Carolina.
- Only **4 of the 13** crew members of both aircraft survived.
- The second plane returning from Turkey contained 4 atomic bombs. And all four were going to fall squarely on Palomares in the east of Almería.
- Of the **four bombs**, three fell on land and could be located. Numerous witnesses assured the US Army detachment to Palomares that the **stray bomb** had landed in the water. But, **how to find it?**

The casings of two B28 nuclear bombs involved in the Palomares incident, on display at the National Atomic Museum, in Albuquerque

Alan Turing and Enigma code

- During the Second World War II, Alan Turing and his team worked on decrypt the **enigma code**.
- Algorithm developed was called **"bamburismus"**. This was used to **decrypt messages sent by the German Navy**.
- The Enigma machine consisted of a keyboard, a panel where the letters lit up and **several rotors**.
- To encrypt a message, the rotors were placed in a **certain position** and the message was written and the encrypted message was displayed on the panel.
- To decrypt an encrypted message, the process was symmetrical. Simply **set the rotors to the initial configuration** and type in the encrypted message, which would appear decoded on the panel. But, **how bamburismus worked to decrypt messages?**

Enigma machine

How these problems were solved?

Palomares Bomb

- **A prior probability was assigned to each area of the map** based on the subjective knowledge of the experts. This probability could be **updated** with incoming information on the search.
- That day, **Paco was fishing 90 metres from the place where the bomb fell**, and was able to act as a guide for the Americans who had been sent to Palomares.
- With his help, **they were able to formulate another map based on the information he gave them**. They finally managed to locate the bomb.

Alan Turing and Enigma code

- As coded messages were received, the belief about the **hypothetical machine configuration was updated**.
- When the **weight of evidence in favour of a particular configuration** of the Enigma machine was sufficiently high, that configuration was considered probable.

Table of Contents

1. A bit of history

UNIVERSITA
POLITÈCNIC
DE VALÈNC

Thomas Bayes (1701-1761)

- He was an **English statistician, philosopher and Presbyterian minister**.
- Specifically, he was the **eldest of 7 children** born to Ann and Joshua Bayes. Joshua was one of the first seven Presbyterian reverends, a branch of the Protestant church that did not support the Anglican church.
- He moved to London, but as a Presbyterian, **Bayes was unable to study theology** in that city, and had to travel to the city of Edinburgh to study theology. It seems that he studied also **Mathematics and Logic**.
- Thomas was interested in studying **the most probable causes** of what was happening in order to prove God's existence and benevolence.

The work that changed everything

- In this respect, one of the people Thomas most admired was **Sir Isaac Newton**. Newton was trying to prove that if **there were such rules, it was because there was a God who organized everything**.
- Bayes devises a procedure by which, **starting from, zero knowledge, one can learn from observations (data, facts) to get to know what causes them**.
- During his lifetime, he did not want to **publish his theorem** because he thought it was irrelevant.
- It only saw the light of day when his friend **Richard Price** retrieved it from a pile of papers. Finally, it appeared in 1763, in the journal Philosophical Transactions under the name: **An Essay toward solving a Problem in the Doctrine of Chances**.

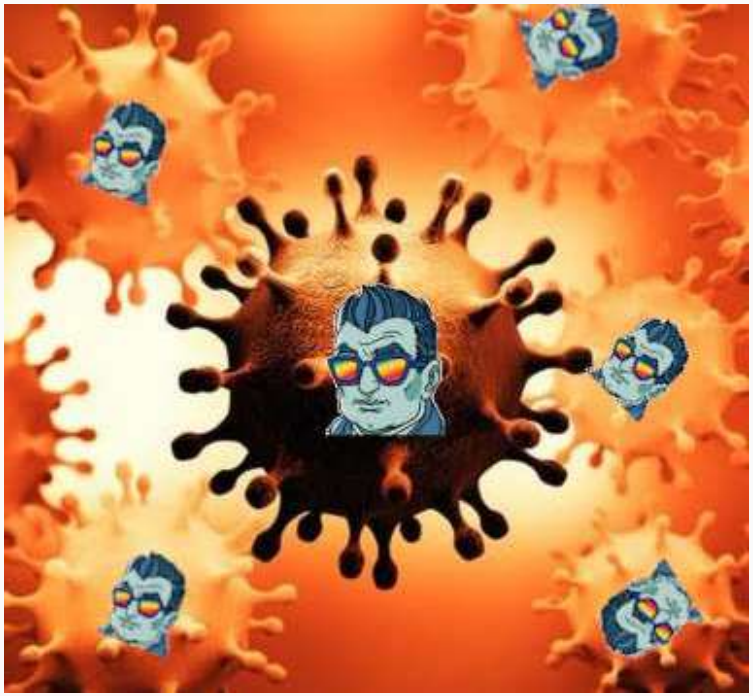
How does it work?

- I know something about the hypothesis -> **Prior distribution**: $Pr(Hypothesis)$.
- I observe the data -> **Likelihood**: $Pr(Data \mid Hypothesis)$.
- I update my knowledge about my hypothesis -> **Posterior distribution**: $Pr(Hypothesis \mid Data)$.

$$Pr(Hypothesis \mid Data) = \frac{Pr(Data \mid Hypothesis) \cdot Pr(Hypothesis)}{Pr(Data)}$$

Example: Bayesianitis

- A test for detecting **Bayesianitis** (a Statistics condition) has been made. This test has:
  - 95 % sensitivity
  - 98 % specificity
- It is known that **1 of 100 statisticians** has this condition. The test is given massively in different universities in Spain. What is the Probability that Laplace, who has **tested + in this test is indeed Bayesianitis +**?
- Laplace has tested positive in the first test, what would be the probability to be **Bayesianitis + if Laplace tests positive again?**



https://elpais.com/ciencia/cafe-y-teoremas/2021-12-07/como-el-ataque-de-pearl-harbor-cambio-la-estadistica-de-las-pruebas-diagnosticas.html

Example: Bayesianitis. Posterior

Hypothesis. Prior information

B is 'true condition' of Laplace:

- B^+: Laplace is Bayesianitis +
- B^-: Laplace is Bayesianitis -

"1/100 prevalence" -> $Pr(B^+) = 0.01$

Data. Observed and then, known

$+_1$ the first test is positive.

95% sensitivity ->
$Pr(+_1 \mid B^+) = 0.95$

98% specificity ->
$Pr(-_1 \mid B^-) = 0.98$

Updated Hypothesis. Posterior information

$$Pr(B^+ \mid +_1) = \frac{Pr(+_1 \mid B^+)Pr(B^+)}{Pr(+_1)}$$

$$= \frac{Pr(+_1 \mid B^+)Pr(B^+)}{Pr(+_1 \mid B^+)Pr(B^+) + Pr(+_1 \mid B^-)Pr(B^-)} =$$

$$= \frac{0.95 \times 0.1}{0.95 \times 0.1 + 0.02 \times 0.9} = 0.324$$

- **32.4% of those Statisticians getting a `(+)` test are Bayesianitis +.**

Example: Bayesianitis. Testing again. Updating

- What would be the probability to be Bayesianitis+ if Laplace tests positive again?

Hypothesis. Prior information

Prior information now is posterior of the previous process.

$$P(B^+ \mid +_1) = 0.324$$

Data. Observed and then, known

$+_2$ the second test is positive.

95% sensitivity ->
$$P(+_2 \mid B^+) = 0.95$$

98% specificity ->
$$P(-_2 \mid B^-) = 0.98$$

Updated Hypothesis. Posterior information

$$P(B^+ \mid +_2, +_1) = \frac{P(+_2 \mid B^+, +_1)P(B^+ \mid +_1)}{P(+_2 \mid +_1)}$$

$$= \frac{P(+_2 \mid B^+)P(B^+ \mid +_1)}{P(+_2 \mid B^+, +_1)P(B^+ \mid +_1) + P(+_2 \mid B^-, +_1)P(B^-}$$

$$= \frac{0.95 \times 0.32}{0.95 \times 0.32 + 0.02 \times 0.68} = 0.958$$

- **95.8% of those Statisticians getting a second `(+)` test are Bayesianitis +.**

Example: Diabetes Screening

- A test for detecting **Diabetes** (a health condition) has been conducted. This test has:
  - 80 % sensitivity
  - 97 % specificity
- It is known that **14.8% of individuals** have this condition. The test is administered widely in different healthcare facilities in Spain. What is the probability that an individual who has **tested + in this test indeed has diabetes +**?
- The individual has tested positive in the first test, what would be the probability of having **diabetes + if they test positive again?**

Example: Diabetes Screening. Posterior

Hypothesis. Prior information

D is 'true condition' of the individual:

- D^+: Individual has diabetes
- D^-: Individual does not have diabetes

"14.8% prevalence" ->
$Pr(D^+) = 0.148$

Data. Observed and then, known

$+_1$ the first test is positive.

80% sensitivity ->
$Pr(+_1 \mid D^+) = 0.80$

97% specificity ->
$Pr(-_1 \mid D^-) = 0.97$

Updated Hypothesis. Posterior information

$$Pr(D^+ \mid +_1) = \frac{Pr(+_1 \mid D^+)Pr(D^+)}{Pr(+_1)}$$

$$= \frac{Pr(+_1 \mid D^+)Pr(D^+)}{Pr(+_1 \mid D^+)Pr(D^+) + Pr(+_1 \mid D^-)Pr(D^-)} =$$

$$= \frac{0.80 \times 0.148}{0.80 \times 0.148 + 0.03 \times 0.852} = 0.822$$

- **82.2% of individuals getting a `(+)` test have diabetes.**
- **What would be the probability for the individual to be diabetic if we conduct a second test?**

Example: Transition to Bayesian Distributions

- We now apply the Bayes theorem to statistical models by changing the notation to use a parameter θ, which represents the true condition:
  - $\theta = 1$ for D^+ (disease present)
  - $\theta = 0$ for D^- (disease absent)
- Test results are recoded as data:
  - $y = 1$ indicates a positive test result -> T^+
  - $y = 0$ indicates a negative test result -> T^-
- **The posterior probability of the individual having the disease given a positive test result is**:

$$p(\theta = 1 \mid y = 1) = \frac{p(y = 1 \mid \theta = 1) \cdot p(\theta = 1)}{p(y = 1 \mid \theta = 1) \cdot p(\theta = 1) + p(y = 1 \mid \theta = 0) \cdot p(\theta = 0)}$$

- $p(\theta = 1)$ represents the prior probability of disease (prevalence), while $p(y = 1 \mid \theta = 1)$ is the likelihood function for a positive test.

2. Bayesian approach

UNIVERSITA
POLITÈCNIC
DE VALÈNC

Density and probability functions

Bayes Theorem

$$Pr(Hypothesis \mid Data) = \frac{Pr(Data \mid Hypothesis) \cdot Pr(Hypothesis)}{Pr(Data)}$$

Bayesian Inference

$$p(\theta \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{\int p(\boldsymbol{\theta}) p(\boldsymbol{y} \mid \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}} \propto p(\boldsymbol{y} \mid \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})$$

- **Likelihood** $p(\boldsymbol{y} \mid \boldsymbol{\theta})$: function that takes information from the data.
- **Prior distribution** $p(\boldsymbol{\theta})$: represents our previous knowledge about the parameter of interest.
- **Posterior distribution** $p(\boldsymbol{\theta} \mid \boldsymbol{y})$: represent what you know after having seen the data. The basis for inference, a distribution, possibly multivariate if more than one parameter.

Bayesian inference. Principles

- Interpretation of probability is related with our **degree of belief** about the event we are considering.
  - For example, in the case of rain, an 80% probability of rain simply tells us that it is more likely to rain than not to rain.
- Information and uncertainty we have about everything unknown is expressible in **terms of probability distributions**.
- It considers as uncertain elements of the problem not only **the data** but also **the parameters**.
- The idea of **repeated sampling is not used** to interpret the properties of estimators.
- **Observed information updates knowledge about the unknown**.
- **Estimators disappear** and are **inferred in terms of probability distributions**.
- **Frequentist approach relies on sample data (present)**. **Bayesian uses also prior information (past)**.

Example: Scoring penalties Valencia C. F.

- 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**.

Example: Scoring penalties Valencia C. F.

Response variable + Data

- $Y = \text{score/miss the penalty}$
- The model is generated by Y
- **Bernoulli** with parameter π, i.e., $Y \sim Ber(\pi)$
- **Likelihood**

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

k: times that a player score a penalty (30).

N: total penalties in 10 seasons (50 matches).

Prior knowledge about the parameter π

- **Beta distribution** seems adecuate to model a proportion π.
- After asking some experts, we end up with a 75 percentage chance to score a penalty.
- We express this uncertainty using percentiles $per_{90} = 0.8$ and $per_{50} = 0.75$.
- The corresponding values for a and b are $a = 83.46$ and $b = 28.05$.
- **Prior distribution**

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

Graphical Model

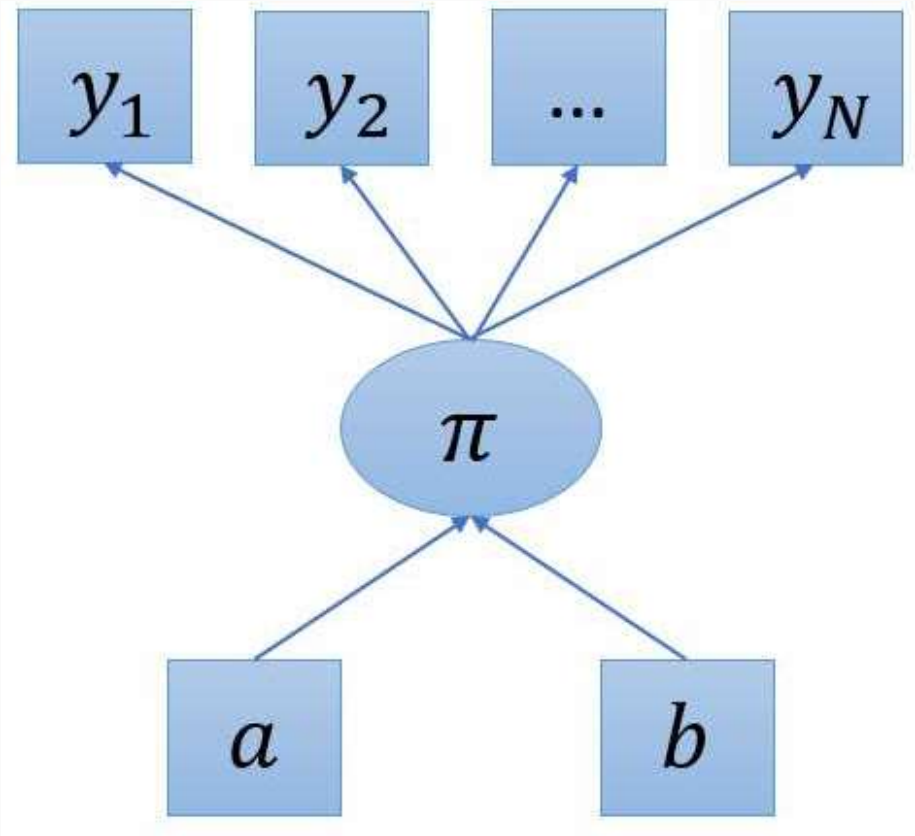
Likelihood

$$p(\boldsymbol{y} \mid \pi) \propto \pi^{k}(1-\pi)^{N-k}$$

Prior distribution

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

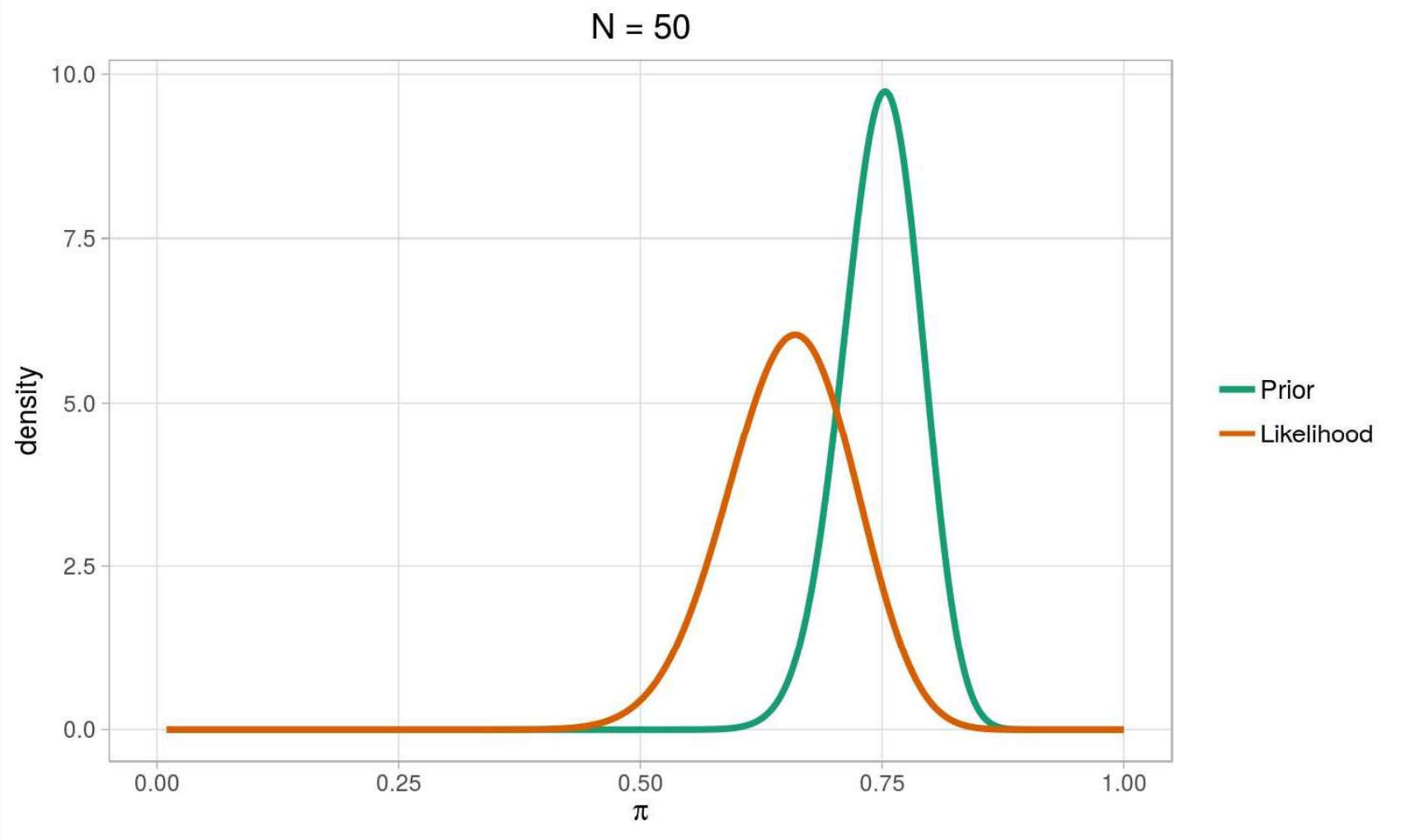
$$\pi \sim \mathrm{Beta}(a, b)$$



Ellipses: variables

Squares: data

Example. Likelihood *vs* Prior



Posterior distribution. Bayesian learning process

Estimating the probability to score a penalty

Likelihood

$$p(\boldsymbol{y} \mid \pi) = \pi^{k}(1-\pi)^{N-k}$$

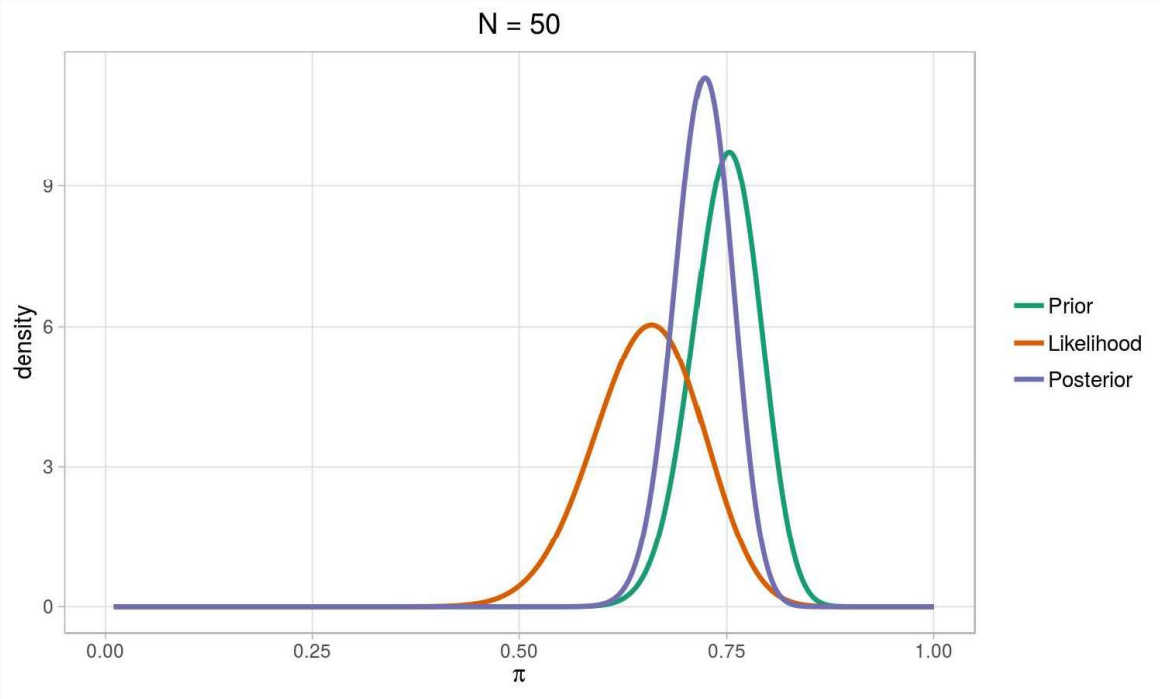
Prior distribution

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

Posterior distribution

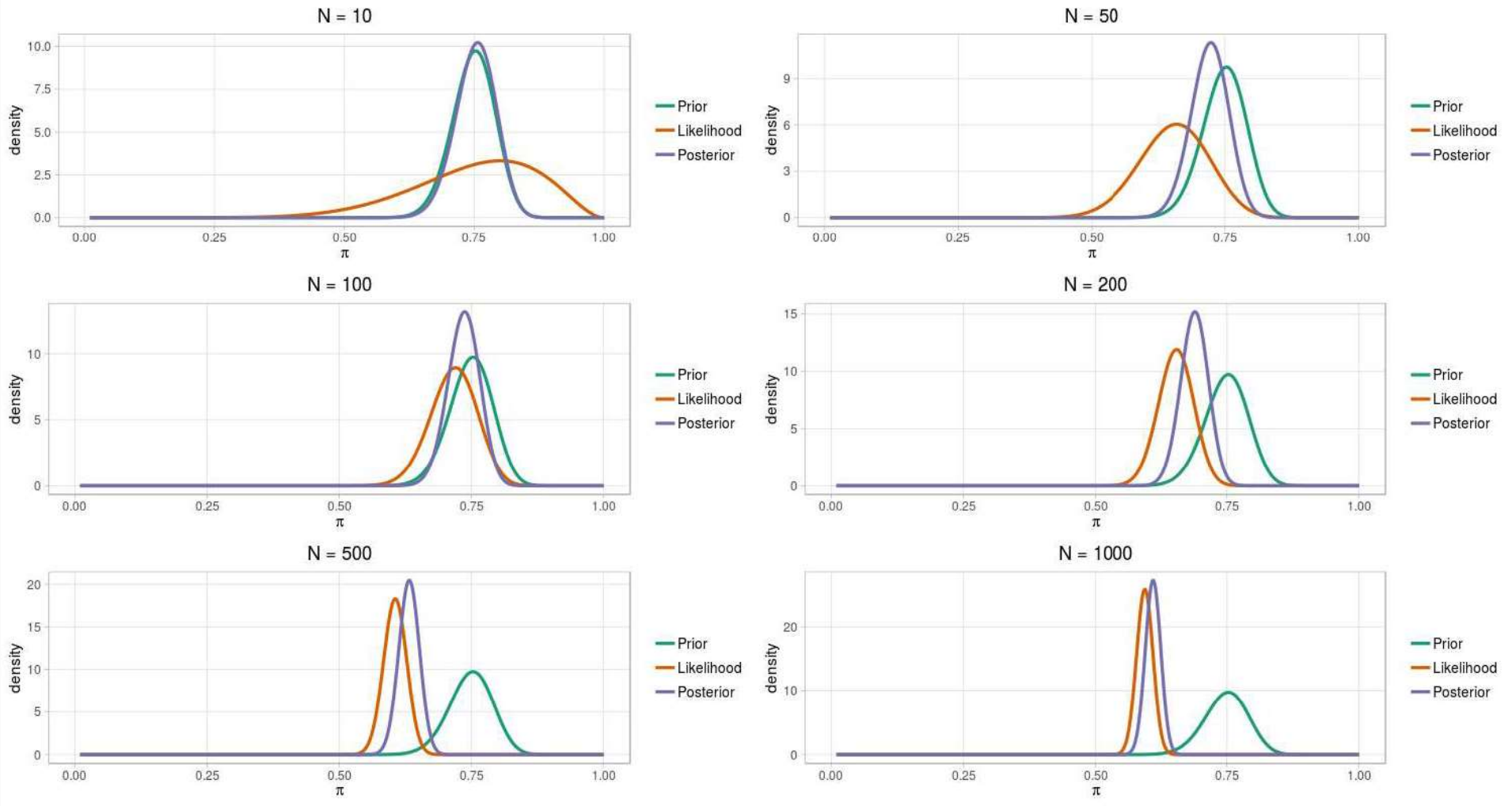
$$\begin{aligned} p(\pi \mid \boldsymbol{y}) &\propto p(\boldsymbol{y} \mid \pi) \cdot p(\pi) \\ &\propto \pi^{k+a-1}(1-\pi)^{N-k+b-1} \end{aligned}$$

$$\pi \mid \boldsymbol{y} \sim \text{Beta}(k+a, N-k+b)$$
$$\pi \mid \boldsymbol{y} \sim \text{Beta}(30+83.46, 20+28.05)$$
$$\pi \mid \boldsymbol{y} \sim \text{Beta}(113.46, 48.05)$$



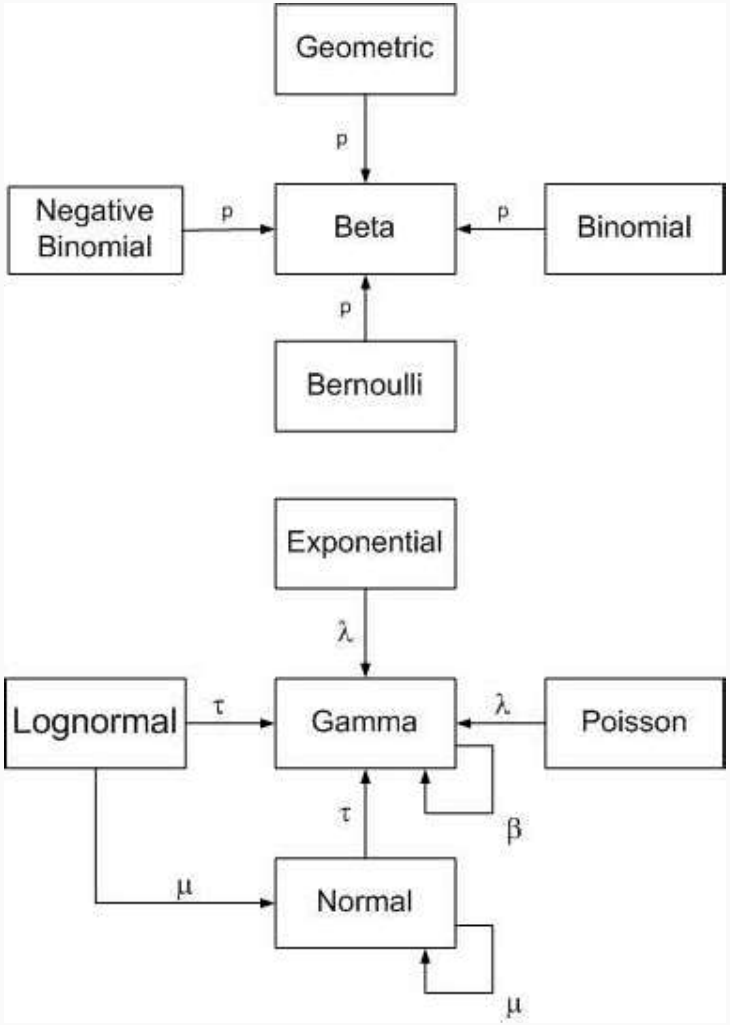
Let's try to understand how a prior works:
https://minaya.shinyapps.io/Beta-Conjugate-Priors/

Data *vs* prior information



Conjugate priors

- **Beta distribution is a conjugate prior** for the Binomial likelihood function.
  - If the posterior distribution $p(\theta \mid y)$ is in the same probability distribution family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called **a conjugate prior for the likelihood function** $p(y \mid \theta)$.
- Compendium Of Conjugate Priors



Describing results: point estimators, credible

- We obtain a **probability or a density function as a posterior**. So, we can deal with the complete distribution

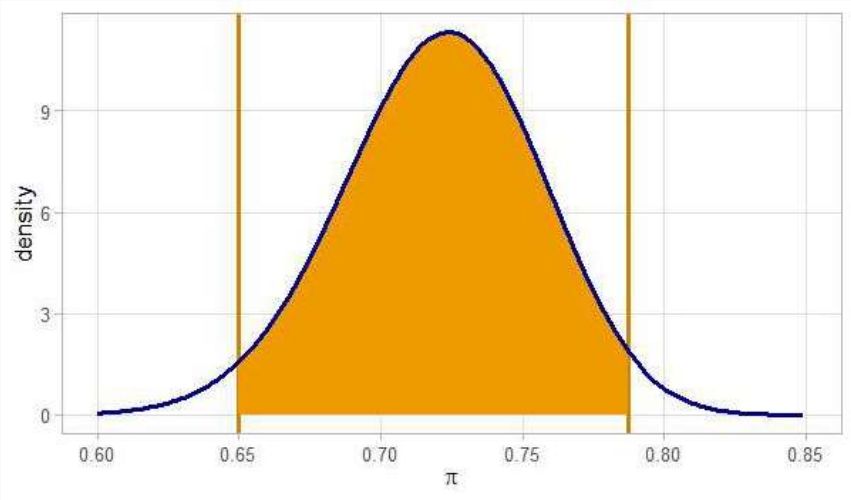
Point estimates

- Mean, median, mode

Credible intervals

- $100(1-\alpha)\%$ credibility interval (CI) for a parameter θ is defined as the pair of values a and b such as :
  $p(\theta \leq a \mid \boldsymbol{y}) = \alpha/2$ and
  $p(\theta \geq b \mid \boldsymbol{y}) = 1 - \alpha/2$

- The $IC_{95\%}(\pi) = (0.65; 0.79)$



Credible interval *vs* Confidence interval

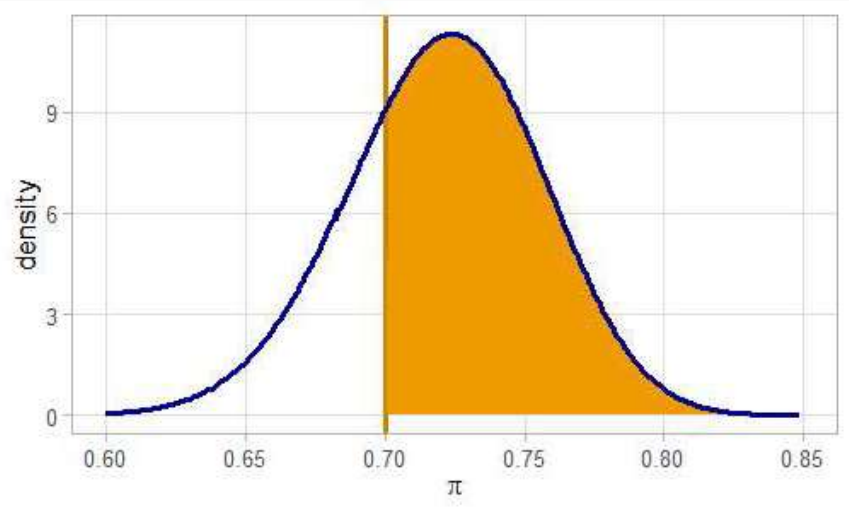
- **Frequentist approach**: a $100(1-\alpha)\%$ confidence interval is defined such that, if the data collection process is repeated again and again, then in the long run, $100(1-\alpha)$ **of the confidence intervals formed would contain the (fixed) unknown parameter value**.

- **Bayesian approach**: a $100(1-\alpha)\%$ credible interval will explicitly indicate **the posterior probability that θ lies within its boundaries**. So, we talk about credibility intervals.

    - The $IC_{95\%}(\pi) = (0.65; 0.79)$ means that the probability for π to be between 0.65 and 0.79 is 0.95.

- Bayesian inference can also provide any probability statements about parameters.

    - For example, we could compute $p(\pi > 0.7 \mid \boldsymbol{y}) = 0.73$.



Hypothesis Testing in Bayesian Framework

- **Established hypotheses**: We want to determine which hypothesis is correct.
  - $H_0 : \theta \in \Theta_0$
  - $H_1 : \theta \in \Theta_1$
- **Solution**:
  - Calculate $\alpha_0 = P(\Theta_0 \mid x)$ and $\alpha_1 = P(\Theta_1 \mid x)$.
  - Reject H_0 if $\alpha_0 < \alpha_1$.
- **Alternative approaches**:
  - **Odds a priori**: π_0/π_1 (where π_i is the prior probability of Θ_i).
  - **Odds a posteriori**: α_0/α_1.
  - **Bayes Factor**: $\frac{\pi_0/\pi_1}{\alpha_0/\alpha_1}$. This measures how much the data favors H_0 over H_1.

With the information available, can we confirm $\pi > 0.9$?

3. Predictions

UNIVERSITA
POLITÈCNIC
DE VALÈNC

Predictions

Prior predictive distribution

- Using just the **previous information** about the population.
- **Before performing the experiment** one can infer about the most/least probable values to be observed.

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

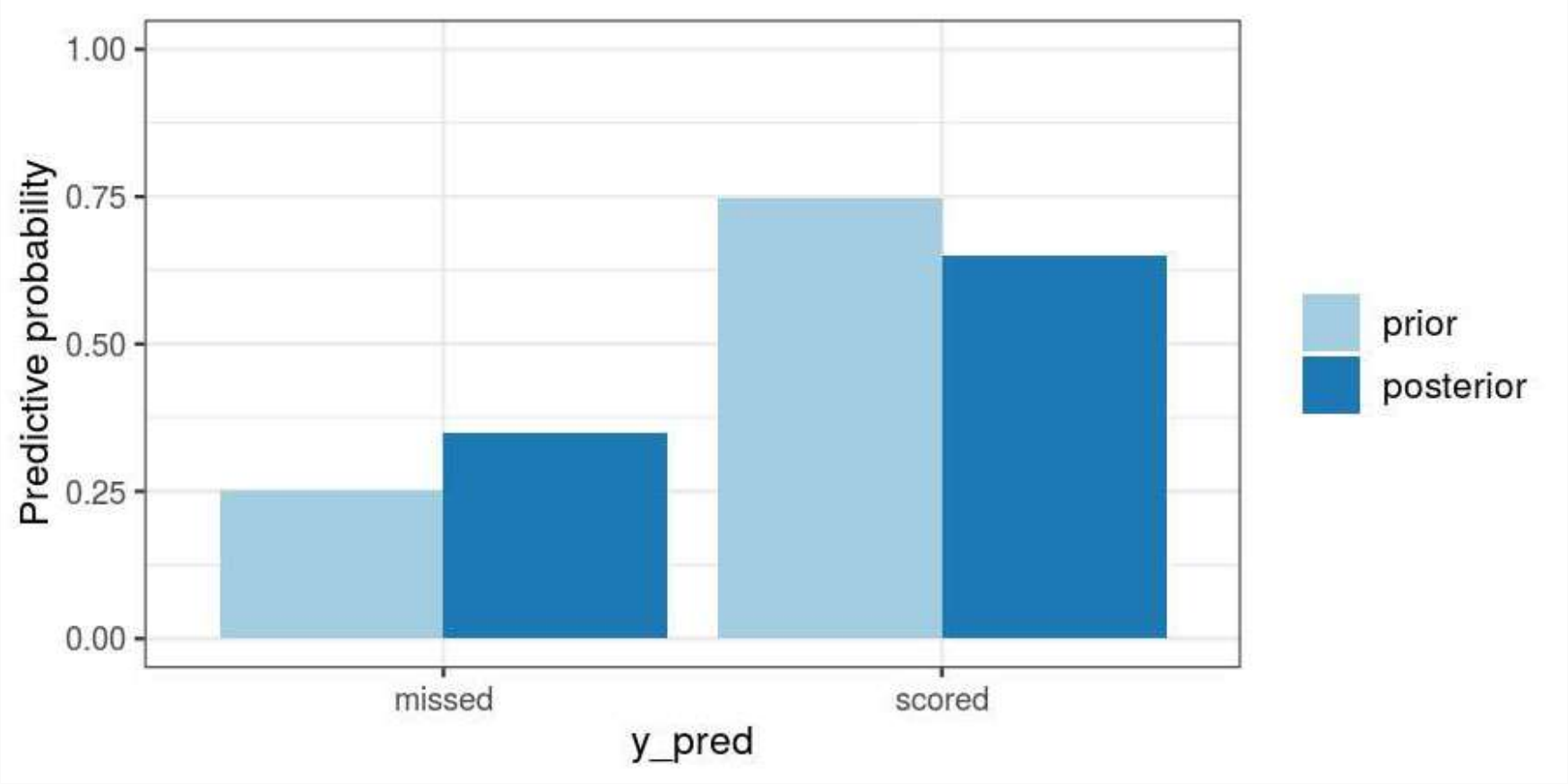
Posterior predictive distribution

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

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

Prior *vs* Posterior predictive

- Rubén Baraja 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.



What we have learned so far

- **ALL uncertainty** is quantified through probability distributions.
- Bayes theorem (probability calculus) is the tool to **combine several sources of uncertainty**.
- **Prior information is totally separated from information from the data** (use the data twice is forbidden!)
- Prior information **is updated by data information into posteriors**.
- **Posterior distributions combine and quantify** the information/uncertainty about the state of interest and are the pillar blocks of Bayesian inference.

But..., in the meantime...

- 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

UNIVERSITA
POLITÈCNIC
DE VALÈNC

Example: Scoring penalties Valencia C. F.

- 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**.

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**?

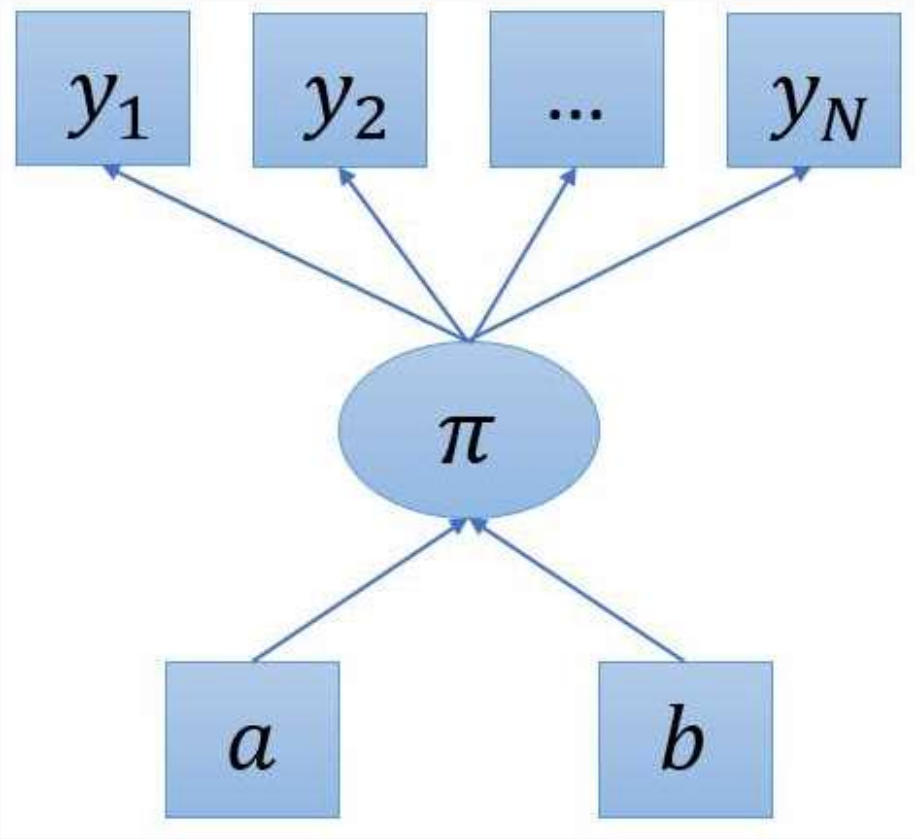
Three ways to do so

1. **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.

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

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



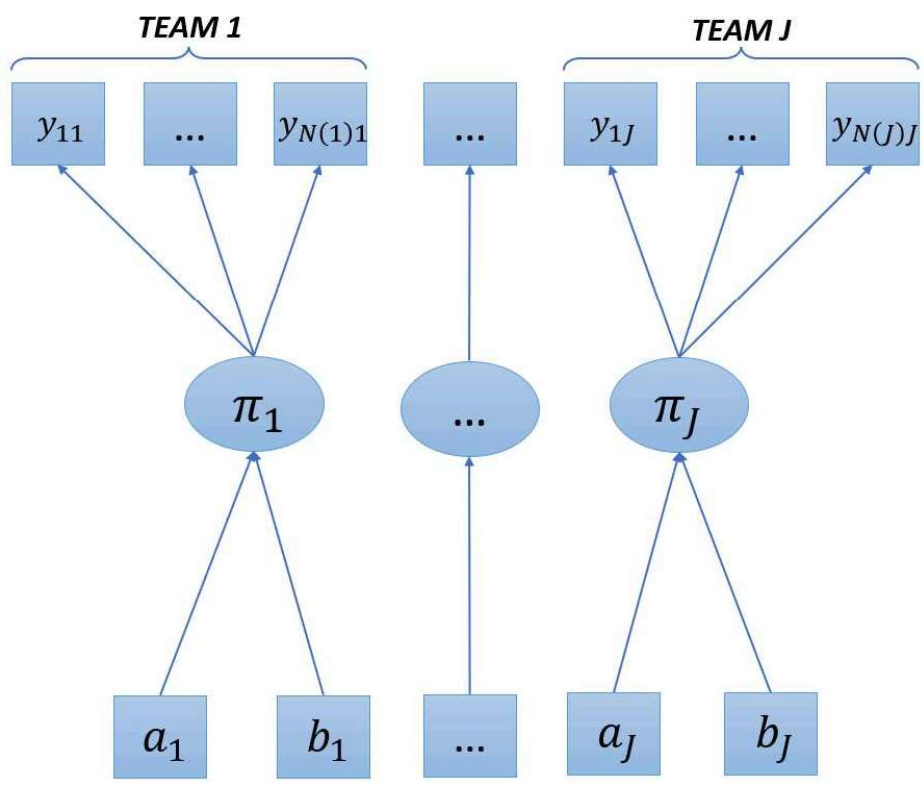
Three ways to do so

2. **Each team is different and has nothing in common with the others.**

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

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

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



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** are at least **similar**.

Three ways to do so

3. Consider a **hierarchical model**.

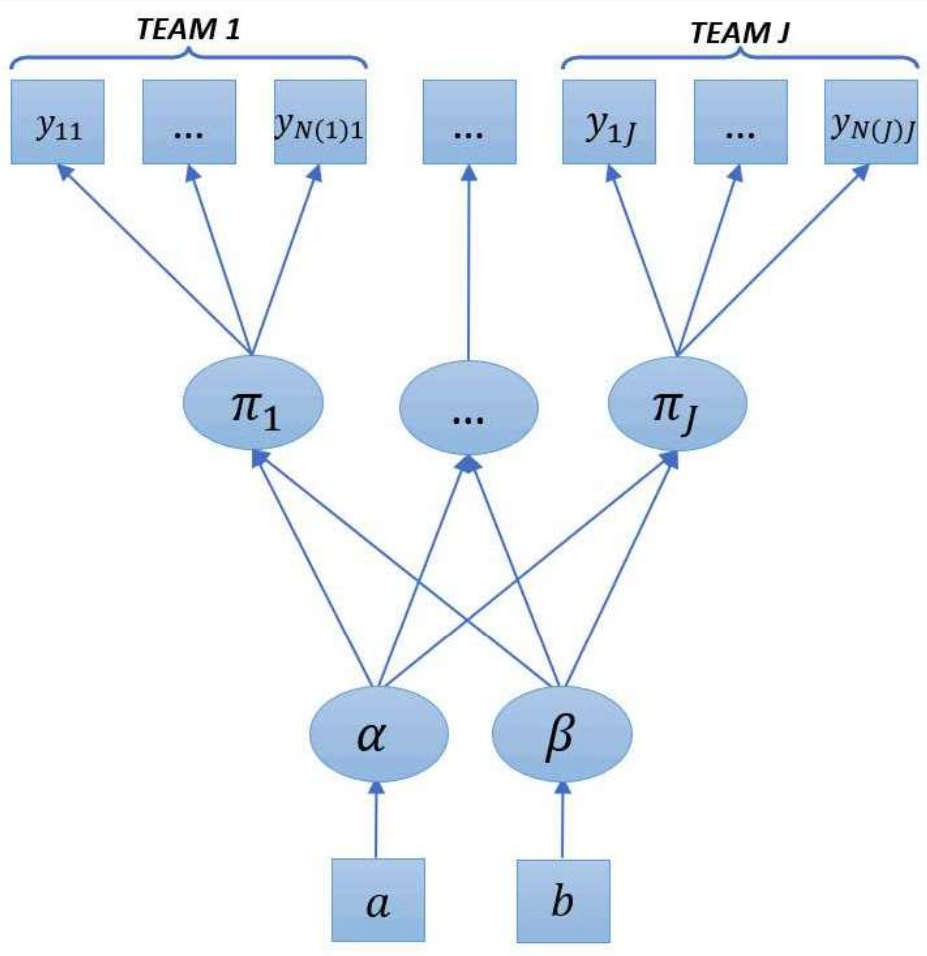
- The parametric vector $\boldsymbol{\pi} = \pi_1, \ldots, \pi_J$ is a **random sampling from a common distribution** that depends on a vector of **hyperparameters**, α and β, partial or totally unknown.

- The model

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

$$\pi_j \sim \text{Beta}(\alpha, \beta)$$

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



Numerical approaches

- When applying Bayesian Statistics, most of the usual models do not yield analytic expressions for neither the posterior nor the predictive posterior 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**.

References

Blogs

- https://towardsdatascience.com/bayesian-updating-simply-explained-c2ed3e563588
- https://medium.com/callisto-media-lab-blog/the-counter-intuitive-stats-principle-that-broke-the-enigma-code-dab6ce69d423
- https://translatingnerd.com/2018/02/08/searching-for-lost-nuclear-bombs-bayes-rule-in-action/

Blogs (Spanish)

- http://anabelforte.com/2020/04/08/thomas-bayes/
- http://anabelforte.com/2020/07/23/un-teorema-para-el-siglo-xxi/
- http://anabelforte.com/2022/04/03/en-bayesiano-como/
- https://picanumeros.wordpress.com/2021/04/18/la-estadistica-detras-del-rescate-de-la-bomba-de-palomares/

References

Books

- McGrayne, S. B. (2011). The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, & Emerged Triumphant from Two Centuries of C. Yale University Press.
- Hoff, P. D. (2009). A first course in Bayesian statistical methods (Vol. 580). New York: Springer.
- Bolstad, W. M., & Curran, J. M. (2016). Introduction to Bayesian statistics. John Wiley & Sons.
- Kruschke, J. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.
- **Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, Donald B. Rubin (2013). Bayesian Data Analysis. Chapman and Hall/CRC**
- **Lesaffre, E., & Lawson, A. B. (2012). Bayesian Biostatistics. Chapman & Hall/CRC Biostatistics Series**
- Broemeling, L. D. (2013). Bayesian Methods in Epidemiology. Chapman and Hall/CRC

ADIM: An introduction to Bayesian inference

Master's Degree in Data Analysis, Process Improvement and Decision Support Engineering

Joaquín Martínez-Minaya, 2024-12-02
VAlencia BAyesian Research Group
Statistical Modeling Ecology Group
Grupo de Ingeniería Estadística Multivariante

jmarmin@eio.upv.es

UNIVERSITA
POLITÈCNIC
DE VALÈNC